REFERENCES

1. Cheung, G.K.M., Kanade, T., Bouguet, J.Y., Holler, M., 2000, "A Real Time System for Robust 3D Voxel Reconstruction of Human Motions", **IEEE Conference on Computer Vision and Pattern Recognition**, 13-15 Jun 2000, South Carolina, pp. 714-720.

2. Kehl, R., Bray, M. and Gool, L. V., 2005, "Full Body Tracking from Multiple Views Using Stochastic Sampling, **IEEE Conference on Computer Vision and Pattern Recognition**, 13-15 Jun 2000, Washington DC, pp. 129-136.

3. Gall, J., Stoll, C., Aguiar, E., Theobalt, C., Rosenhahn, B. and Seidel, H., 2009, "Motion Capture Using Joint Skeleton Tracking and Surface Estimation", **IEEE Conference on Computer Vision and Pattern Recognition**, 20-25 Jun 2009.

4. Bottino, A. and Laurentini, 2001, "A Silhouette-based Technique for the Reconstruction of Human Movement", **Computer Vision and Image Understanding**, Vol. 83, No. 1, pp. 7995.

5. Kelly, P.H., Katkere, A., Kuramura, D.Y., Moezzi, S., Chatterjee, S. and Jain, R., 1995, "An Architecture for Multiple Perspective Interactive Video", **ACM Conference on Multimedia**, pp. 201-212.

6. Jain, R and Wakimoto, K., 1995, "Multiple Perspective Interactive Video", **International Conference on Multimedia Computing and Systems**, 15-18 May 1995, Washington DC, pp. 201-211.

7. Kehl, R. and Gool, L.V., 2006, "Markerless Tracking of Complex Human Motions from Multiple Views", **Computer Vision and Image Understanding**, Vol. 104, No. 2-3, pp. 190-209.

8. Shakhnarovich, K., Viola, P.A. and Darrell, T., 2003, "Fast Pose Estimation with Parameter-sensitive Hashing", **The International Conference on Computer Vision**, France, pp. 750-759.

9. Mori, G. and Malik, J., 2006, "Recovering 3D Human Body Configurations Using Shape Contexts", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 28, pp. 1052-1062.

10. Howe, N.R., 2007, "Silhouette Lookup for Monocular 3D Pose Tracking", **Journal Image Vision Computing**, Vol. 25, pp. 331-341.

11. Brand, M., 1999, "Shadow Puppetry", **The International Conference on Computer Vision**, Sep 1999, Kerkyra, Greece, pp. 1237-1244.

12. Guo, F. and Qian, G., 2006, "Learning and Inference of 3D Human Poses from Gaussian Mixture Modeled Silhouettes", **The International Conference on Computer Vision**, Hong Kong, pp. 43-47.

13. Agarwal, A. and Triggs, B., 2004, "3D Human Pose from Silhouettes by Relevance Vector Regression", **IEEE Conference on Computer Vision and Pattern Recognition** , 27 June-2July 2004, Washington D.C, pp. 882-888.

14. Howe, N.R., Leventon, M.E. and Freeman, W.T., 2000, "Bayesian Reconstruction of 3D Human Motion from Single-camera Video", **Advances in Neural Information Processing Systems**, Nov 2000, Denver, CO, pp. 820-826.

15. Grauman, K., Shakhnarovich, G. and Darrell, T., 2003, "Inferring 3D Structure with a Statistical Image-based Shape Model", **The International Conference on Computer Vision**, France, pp. 641-647.

16. Ning, H., Xu, W., Gong, Y. and Huang, T., 2008, "Discriminative Learning of Visual Words for 3D Human Pose Estimation", **IEEE Conference on Computer Vision and Pattern Recognition**, 23-28 June 2008, Anchorage, AK, pp. 1-8.

17. Ning, H., Hu, Y and Thomas, H., 2008, "Efficient initialization of Mixtures of Experts for Human Pose Estimation", **IEEE International Confernce on Image Processing**, 12-15 Oct 2008, San Diego, CA, USA, pp. 2164-2167.

18. Taylor, C. J., 2000, "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image", **IEEE Conference on Computer Vision and Pattern Recognition**, 13-15 Jun 2000, South Carolinap, pp. 677-684.

19. Mori, G. and Malik, J., 2002, "Estimating Human Body Configurations Using Shape Context Matching", **European Conference on Computer Vision**.

20. Rius, I., Rowe, D., Gonzlez, J. and Roca, F.X., 2005, "3D Action Modeling and Reconstruction for 2D Human Body Tracking", **Pattern Recognition and Image Analysis**, Vol. 3687, pp. 146-154.

21. Remondino, F. and Roditakis, A., 2003, "3D Reconstruction of Human Skeleton from Single Images or Monocular Video Sequence", **DAGM-Symposium**, pp. 100-107.

22. Lao, W., Han, J. and With, P.H.N. de, 2006, "3D Modeling for Capturing Human Motion from Monocular Video", **Symposium on Information Theory in the Benelux**, pp. 299-306.

23. Roodsarabi, N. and Behrad, A., 2008, "3D Human Motion Reconstruction Using Video Processing", **Image and Signal Processing**, pp. 386-395.

24. Gao, J. and Shi, J., 2004, "Multiple Frame Motion Inference Using Belief Propagation", **IEEE International Conference on Automatic Face and Gesture Recognition**, May 2004, pp. 875- 880.

25. Remondino, F. and Roditakis, A., 2003, "Human Figure Reconstruction and Modeling from Single Image or Monocular Video Sequence", **Fourth International Conference on 3D Digital Imaging and Modeling**, pp. 116-123.

26. Sminchisescu, C. and Jepson, A., 2004, "Generative Modeling for Continuous Non-Linearly Embedded Visual Inference", **International Conference on Machine Learning**, Banff, pp. 759766.

27. Sminchisescu, C. and Jepson, A., 2004, "Variational Mixture Smoothing for Non-Linear Dynamical Systems", **IEEE International Conference on Computer Vision and Pattern Recognition**, Washington D.C., pp. 608615.

28. Shen, C., Hengel, A.V., Dich, A. and Brooks, M.J., 2004, "2D Articulated Tracking with Dynamic Bayesian Networks", **IEEE International Conference on Computer and Information Technology**, Sept 2004, pp.130-136.

29. Park, M., Liu, Y. and Collins, R.T., 2008, "Efficient Mean Shift Propagation for Vision Tracking", **IEEE Conference on Computer Vision and Pattern Recognition**, 23-28 Jun 2008 , Anchorage, AK, pp.1-8.

30. Poppe, R., 2007, "Vision-based Human Motion Analysis: an Overview", **Computer Vision and Image Understanding**, Vol. 108, pp. 4-18.

31. Moeslund, T.B., Hilton, A. and Kruger, V., 2006, "A Survey of Advances in Vision-based Human Motion Capture and Analysis", **Computer Vision and Image Understanding**, Vol. 104, pp. 90-126.

32. Moeslund, T.B. and Granum, E., 2001, "A Survey of Computer Vision-Based Human Motion Capture", **Computer Vision and Image Understanding**, Vol. 81, pp. 231-268.

33. Aggarwal, J.K. and Cai, Q., 1999, "Human Motion Analysis: a Review", **Computer Vision and Image Understanding**, Vol. 73, No. 3, pp. 428440.

34. Gavrila, D.M., 1999, "The Visual Analysis of Human Movement: a Survey", **Computer Vision and Image Understanding**, Vol. 73, No. 1, pp. 8292.

35. Wang, J.J. and Singh, S., 2003, "Video Analysis of Human Dynamics: a Survey", **Real-Time Imaging**, Vol. 9, No. 5, pp. 321346.

36. Rourke, J. and Badler, N.I., 1980, "Model-based Image Analysis of Human Motion Using Constraint Propagation", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 2, No. 6, pp. 522-536.

37. Hogg, D., 1983, "Model-based Vision: a Program to See a Walking Person", **Image and Vision Computing**, Vol. 1, No.1, pp. 5-20.

38. Rohr, K., 1994, "Towards Model-based Recognition of Human Movements in Image Sequences", **Computer Vision, Graphics, and Image Processing: Image Understanding**, Vol. 59, No. 1, pp. 94-115.

39. Sidenbladh, H., Black, M.J. and Fleet, D.J., 2000, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion", **The European Conference on Computer Vision**, June 2000, Dublin, Ireland, pp. 702-718.

40. Delamarre, Q. and Faugeras, O., 2001, "3D Articulated Models and Multiview Tracking with Physical Forces", **Computer Vision and Image Understanding**, Vol. 81, No. 3, pp. 328-357.

41. Gavrila, D. and Davis, L., 1996, "3-D Model-based Tracking of Humans in Action: A Multi-view Approach", **IEEE Conference on Computer Vision and Pattern Recognition**, San Francisco, pp. 73-80.

42. Felzenszwalb, P.F. and Huttenlocher, D.F., 2005, "Pictorial Structures for Object Recognition", **International Journal of Computer Vision**, Vol. 61, No. 1, pp. 55-79.

43. Sigal, L., Isard, M., Sigelman, B., and Black, M.J., 2005, "Attractive People: Assembling Loose-limbed Models Using Nonparametric Belief Propagation", **Advances in Neural Information Processing Systems**, Vancouver, Canada, pp. 1539-1546.

44. Pavlovic, V.I., Rehg, J.M., Cham, T., and Murphy, K.P., 1999, "A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamic Models, **International Conference on Computer Vision**, Kerkyra, September 1999, Greece, pp. 94101.

45. Agarwal, A. and Triggs, B., 2004, "Tracking Articulated Motion Using a Mixture of Autoregressive Models", **The European Conference on Computer Vision**, May 2004, Prague, Czech Republic, pp. 54-65.

46. Deutscher, J. and Reid, I., 2005, "Articulated Body Motion Capture by Stochastic Search", **International Journal of Computer Vision**, Vol. 61, No. 2, pp. 185205.

47. Sudderth, E., Mandel, M.I., Freeman, W.T. and Willsky, A.S., 2004, "Distribution Occlusion Reasoning for Tracking with Nonparametric Belief Propagation", **Eighteenth Annual Conference on Neural Information Processing Systems**, 13-18 Dec 2004, Vancouver, Canada.

48. Blake, J.D.A. and Reid, I., 2000, "Articulated Body Motion Capture by Annealed Particle Filtering", **IEEE Conference on Computer Vision and Pattern Recognition**, 13-15 Jun 2000, SC, USA, pp. 126-133, 2000.

49. Gritai, A., Basharat, A. and Shah, M., 2008, "Geometric Constraints on 2D Action Models for Tracking Human Body", **International Conference on Pattern Recognition**, 8-11 Dec. 2008 , Florida, pp. 1-4.

50. Lee, M.W., Cohen, I. and Jung, S.K., 2002, "Particle Filter with Analytical Inference for Human Body Tracking", **Workshop on Motion and Video Computing**, Dec 2002, Orlando, FL, pp. 159-168.

51. McKenna, S. J., Raja, Y. and Gong, S., 1999, "Tracking Color Objects Using Adaptive Mixture Models", **Image and Vision Computing**, Vol. 17, No. 3-4, pp.225-231.

52. Wren, C.R., Azarbayejani, A., Darrell, T. and Pentland, A.P., 1996, "Pfinder: Real-time Tracking of the Human Body", **IEEE International Conference on Automatic Face and Gesture Recognition**, 14-16 Oct 1996, New England, pp. 51-56.

53. Ramanan, D. and Forsyth, D.A., 2003, "Finding and Tracking People from the Bottom Up", **IEEE Conference on Computer Vision and Pattern Recognition**, June 2003, Madison, WI, pp. 467-474.

54. Khongkraphan, K. and Kaewtrakulpong, P., 2007, "Robust Contour Tracking in Cluttered Background Using Snake on Weighted-gradient Image", **International Workshop on Advanced Image Technology**, 8-9 Jan 2007, Bangkok.

55. Agarwal, A. and Triggs, B., 2006, "A Local Basis Representation for Estimating Human Pose from Cluttered Images", **The Asian Conference on Computer Vision**, India, pp. 50-59.

56. Bregler, C., 1997, "Learning and Recognizing Human Dynamics in Video Sequences", **IEEE Conference on Computer Vision and Pattern Recognition**, 17-19 Jun 1997, San Juan, Puerto Rico, pp. 568-574.

57. Sminchisescu, C. and Triggs, B., 2003, "Estimating Articulated Human Motion with Covariance Scaled Sampling", **International Journal of Robotic Research**, Vol. 22, No. 6, pp. 371392.

58. Cheung, K.M.G., Baker, S. and Kanade, T., 2003, "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture", **IEEE Conference on Computer Vision and Pattern Recognition**,18-20 June 2003, pp. 77-84.

59. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M., 2004, "Tracking Loose-limbed People", **IEEE International Conference on Computer Vision and Pattern Recognition**, Jun 2004, Washington, DC ,pp. 421-428.

60. , Zhao, X. and Liu, Y., 2007, "Generative Estimation of 3D Human Pose Using Shape Contexts Matching", **The Asian Conference on Computer Vision**, pp. 419-429.

61. Barron, C. and Kakadiaris, I.A., 2001, "Estimating Anthropometry and Pose from a Single Uncalibrated Image", **Computer Vision and Image Understanding**, Vol. 81, No. 3, pp. 269284.

62. Isard, M.A. and Blake, A., 1998, "CONDENSATION Conditional Density Propagation for Visual Tracking", **International Journal of Computer Vision**, Vol. 29, No. 1, pp. 5-28.

63. Agarwal, A. and Triggs, B., 2004, "3D Human Pose from Silhouettes by Relevance Vector Regression", **IEEE Conference on Computer Vision and Pattern Recognition** , 27 June-2July 2004, Washington D.C, pp. 882-888.

64. Ramanan, D. and Sminchisescu, C., 2006, "Training Deformable Models for Localization", **IEEE International Conference on Computer Vision and Pattern Recognition**, 17-22 June 2006, pp. 206-213.

65. Cohen, I. and Lee, M.W., 2002, "3D Body Reconstruction for Immersive Interaction", **Second international Workshop on Articulated Motion and Deformable Objects**, Spain, pp. 21-23.

66. Sidenbladh, H., Black, M.J. and Sigal, L., 2002, "Implicit Probabilistic Models of Human Motion for Synthesis and Tracking", **European Conference on Computer Vision**, May 2002, Copenhagen, Denmark, pp. 784800.

67. Zhou, Z., Prugel-Bennett, A and Damper, R.I., "A Bayesian Framework for Extracting Human Gait Using Strong Prior Knowledge", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 28, No.11, pp. 1738-1752.

68. Agarwal, A. and Triggs, B., 2004, "3D Human Pose from Silhouettes by Relevance Vector Regression", **IEEE Conference on Computer Vision and Pattern Recognition** , 27 June-2July 2004, Washington D.C, pp. 882-888.

69. Sminchisescu, C. and Triggs, B., 2003, "Estimating Articulated Human Motion with Covariance Scaled Sampling", **International Journal of Robotic Research**, Vol. 22, No. 6, pp. 371392.

70. Ning, H., Tan, T., Wang, L. and Hu, W., 2004, "People Tracking Based on Motion Model and Motion Constraints with Automatic Initialization", **Journal of Pattern Recognition**, Vol. 37, No. 7, pp. 1423-1440.

71. Sidenbladh, H., 2004, "Detecting Human Motion with Support Vector Machines", **International Conference on Pattern Recognition**, 23-26 Aug 2004, pp. 188-191.

72. Choo, K. and Fleet, D.J., 2001, "People Tracking Using Hybrid Monte Carlo Filtering", **IEEE Conference on Computer Vision and Pattern Recognition**, 5-6 Dec 2002, pp. 321-328.

73. Chang, C. and Ansari, R., 2003, "Kernel Particle Filter: Iterative Sampling for Efficient Visual Tracking", **International Conference on Image Processing**, 14-17 Sep 2003, pp.977-980.

74. Navaratnam, R., Thayananthan, A. Torr, P.H. and Cipolla, R., 2006, "Hierarchical Part-based Human Body Pose Estimation", September 2005, Oxford, **BMVCThe British Machine Vision Conference**.

75. Ju, S.X., Black, M.J. and Yacoob, Y., 1996, "Cardboard People: a Parameterized Model of Articulated Image Motion", **International Conference on Automatic Face and Gesture Recognition**, Killington, VT, pp. 38-44.

76. Heisele, B. and Wohler, C., 1998, "Motion-based Recognition of Pedestrians", **International Conference on Pattern Recognition**, 16-20 Aug 1998, Brisbane, Australia, pp. 1325-1330.

77. Bregler, C., 1997, "Learning and Recognizing Human Dynamics in Video Sequences", **IEEE Conference on Computer Vision and Pattern Recognition**, 17-19 Jun 1997, San Juan, Puerto Rico, pp. 568-574.

78. Wren, C.R., Clarkson, B.P. and Pentland, A.P., 2000, "Understanding Purposeful Human Motion", **International Conference on Automatic Face and Gesture Recognition**, 28-30 Mar 2000, Grenoble, France, March, pp. 378-383.

79. Sato, K., Maeda, T., Kato, H. and Inokuchi, S., 1994, "CAD-based Object Tracking with Distributed Monocular Camera for Security Monitoring", **CAD-Based Vision Workshop, Champion**, Feburary 1994, PA, pp. 291-297.

80. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M., 2004, "Tracking Loose-limbed People", **IEEE International Conference on Computer Vision and Pattern Recognition**, Jun 2004, Washington, DC ,pp. 421-428.

81. Hua, G., Yang, M.H. and Wu, Y., 2005, "Learning to Estimate Human Pose with Data Driven Belief Propagation", **IEEE Conference on Computer Vision and Pattern Recognition**, 20-25 Jun 2005, pp. 747-754.

82. Ramanan, D., Forsyth, D. and Barnard, K., 2005, "Detecting, Localizing, and Recovering Kinematics of Textured Animals", **IEEE International Conference on Computer Vision and Pattern Recognition**, Jun 2005, pp. 635-642.

83. Agarwal, A. and Triggs, B., 2004, "Learning to Track 3D Human Motion from Silhouettes", **International Conference on Machine Learning**, pp. 9-16.

84. Sminchisescu, C., Kanaujia, A., Li, Z. and Metaxas, D.N., 2005, "Discriminative Density Propagation for 3D Human Motion Estimation", **IEEE International Conference on Computer Vision and Pattern Recognition**, Jun 2005, San Diego, CA, pp. 390-397.

85. Fukunaga, K. and Hostetler, L., 1975, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", **IEEE Transactions on Information Theory**, Vol. 21, No. 1, pp. 32-40.

86. Yizong, C., 1995, "Mean Shift, Mode Seeking, and Clustering", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 17, No. 8, pp.790-799.

87. Vedaldi, A. and Soatto, S., 2008, "Quick Shift and Kernel Methods for Mode Seeking", **European Conference on Computer Vision**, Aug 2008.

88. Zheng, H., Daoudi, M. and Jedynak, B., 2004, "From Maximum Entropy to Belief Propagation: an Application to Skin Detection", **The British Machine Vision Conference**.

89. Gupta, M. D., Rajaram, S., Petrovic, N. and Huang, T. S., 2005, "Non-parametric Image Super-resolution Using Multiple Images", **IEEE International Conference on Image Processing**, pp. 8992.

90. Pappas, T. N. and Jayant, N.S., 1988, "An Adaptive Clustering Algorithm For Image Segmentation", **International Conference on Computer Vision**, 5-8 Dec 1988, pp. 310-315.

91. Bo, S., Ma, Y. and Zhu, C., 2007, "Image Segmentation by Nonparametric Color Clustering", **International Conference on Computational Science**, 27-30 May 2007, Beijing, China, pp. 898-901.

92. Felzenszwalb, P. F. and Huttenlocher, D. P., 2004, "Efficient Belief Propagation for Early Vision", **IEEE International Conference on Computer Vision and Pattern Recognition**, Los Alamitos, CA, USA, pp. 261268.

93. Felzenszwalb, P. F. and Huttenlocher, D. P., 2006, "Efficient Belief Propagation for Early Vision", **International Journal of Computer Vision**, Vol. 70, No. 1, pp. 4154.

94. Yedidia, J. , Freeman, W. and Weiss, Y., 2001, "Understanding Belief Propagation and its Generalizations", **International Joint Conference on Artificial Intelligence**.

95. Freeman, W. T., Jones, T. R. and Pasztor, E. C., 2002, "Example-based Superresolution", **IEEE Computer Graphics and Applications**, pp. 5665.

96. Gupta, M. D., Rajaram, S., Petrovic, N. and Huang, T. S., 2005, "Non-parametric Image Super-resolution Using Multiple Images", **IEEE International Conference on Image Processing**, pp. 8992.

97. Zheng, H., Daoudi, M. and Jedynak, B., 2004, "From Maximum Entropy to Belief Propagation: an Application to Skin Detection", **The British Machine Vision Conference**.

98. Yizong, C., 1995, "Mean Shift, Mode Seeking, and Clustering", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 17, No. 8, pp.790-799.

99. Sheikh, Y.A., Khan, E.A. and Kanade, T., 2007, "Mode-seeking by medoidshifts", **IEEE International Conference on Computer Vision**, 14-21 Oct 2007, Rio de Janeiro , pp. 1-8.

100. Sudderth, E.B., Ihler, A.T., Freeman, W.T. and Willky, A.S., 2003, "Nonparametric Belief Propagation", **IEEE International Conference on Computer Vision and Pattern Recognition**.

101. Isard, M., 2003, "Pampas: Real-valued Graphical Models for Computer Vision", **IEEE International Conference on Computer Vision and Pattern Recognition**, Jun 2003, pp. 613-620.

102. Sudderth, E.B., Mandel, M.I., Freeman, W.T. and Willsky, A.S., 2004, "Visual Hand Tracking Using Nonparametric Belief Propagation", **IEEE International Conference on Computer Vision and Pattern Recognition**, June 2004.

103. Sudderth, E. B., Mandel, M. I., Freeman, W. T. and Willsky, A. S., 2005, "Distributed Occlusion Reasoning for Tracking with Nonparametric Belief Propagation", **Advances in Neural Information Processing Systems 17**, pp. 13691376.

104. Sigal, L. and Black, M.J., 2006, "Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation", **IEEE International Conference on Computer Vision and Pattern Recognition**, pp. 2041-2048.

105. Freeman, W. T., 2000, "Learning Low-level Vision", **International Journal of Computer Vision**, Vol. 40, No. 1.

106. Weiss, Y., 2000, "Correctness of Local Probability Propagation in Graphical Models with Loops", **Neural Computation**, Vol. 12, No, 1.

107. Wang, H. and Li, H, 2003, "A Spectral clustering Approach to Motion Segmentation based on Motion Trajectory", **IEEE International Conference on Multimedia & Expo**, pp. 793-796.

108. Han, T.X., Ning, H. and Huang, T.S., 2006, "Efficient Nonparametric Belief Propagation with Application to Articulated Body Tracking", **IEEE Conference on Computer Vision and Pattern Recognition**, June 2006, New York, NY, pp. 214-221.

109. Peng, X., Zou, B., Chen, S. and Luo, P., 2009, "Reconstruction of 3D Human Motion Pose from Uncalibrated Monocular Video Sequence", **Informational Journal of Information and Systems Sciences**, Vol. 5, p.503-515.

110. Zou, B., Chen, S., Shi, C., and Providence, U.M., 2009, "Automatic Reconstruction of 3D Human Motion Pose from Uncalibrated Monocular Video Sequences Based on Markerless Human Motion Tracking", **Journal of Pattern Recognition**, Vol. 42, pp. 1559-1571.

111. Wu, Y., Hua, G. and Yu, T., 2003, "Tracking Articulated Body by Dynamic Markov Network", **IEEE International Conference on Computer Vision**, 13-16 Oct 2003, Nice, France, pp.1094-1101.

112. Stuart, G. and Donald, G., 1984, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 6, No. 6, pp. 721-741.

113. Ramanan, D., Forsyth, D.A. and Zisserman, A., 2005, "Strike a Pose:Tracking People by Finding Stylized Poses", **IEEE Conference on Computer Vision and Pattern Recognition**, Jun 2005, pp. 271-278.

114. Khongkraphan, K. and Kaewtrakulpong, P., 2010, "A Novel Method of 2D Articulated Body Tracking under Self-occlusion and Ambiguity", **IEICE Electronics Express (ELEX)**, Vol. 7, No. 15, pp.1106-1111.

115. Cox, I.J. and Hingorani, S.L., 1996, "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and its Evaluation for the Purpose of Virtual Tracking", **IEEE Trans. Pattern Analysis and Machine Intelligence**, Vol. 18, pp. 138-150.

116. Jianhui, Z., Li, L. and Keong, K.C., 2005, "3D Posture Reconstruction and Human Animation from 2D Feature Points", **Computer Graphics Forum**, Vol. 24, No. 4, pp. 759-771.

117. Yang, F. and Yuan, X., 2005, "Human Movement Reconstruction from Video Shot by a Single Stationary Camera, **Journal of Annals of Biomedical Engineering**, Vol. 33, pp.674-684.

118. Zhang, L. and Li, L., 2006, "Human Animation from 2D Correspondence Based on Motion Trend Prediction Source", **The 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases**, Madrid, pp.458-463.

119. **Mocap Dataset** [Online], Avaiable : http://mocap.cs.cmu.edu/motcat.php [2009, March 10].

APPENDIX A
BIG-O ANALYSIS

1. A generative approach (top-down approach)
The main concept of a generative approach is an effort to fit a predefined human model into image data.

Pseudo code of a generative approach

```
For i_1 = 1 to N
    For i_2 = 1 to N
        ⋮
            For i_m = 1 to N
                Measure Similarity by an image observation
            End
    End
End
```

Big-O of a generative approach is $O(N^m)$ where N is the number of samples for each body part and m is the number of the body parts.

2. A part-based approach
The main concept is considering each body part separately and the solution of each part is combined into the global solution. The algorithm will first send messages from all leaf nodes to its adjacent node and then will continue sending messages up to the root.

Pseudo code of a part-based approach for each segment in one iteration

```
For k = 1 to m
    For i = 1 to N (samples of node k)
        For j = 1 to N (sample of parent node k)
            Compute messages (both from i to j and from j to i)
        End
        Compute Marginal probability
    End
End
```

Big-O of a part-based approach is $O(mN^2T)$ [29, 108] where N is the number of

samples for each body part, m is the number of the body parts and T is the number of iterations needed for convergence.

3. Optimal solution by MHT concept

In our approach, each joint can have upto two possible solutions. There are 2^m possible configurations where m is the number of the body parts.

Optimal solution by MHT concept

```
For i = 1 to k
    For j = 1 to 2^m
        Compute smoothness function
    End
End
```

After smoothness computation, such smoothness values are ordered for selection of the k solutions. *Big-O* of optimal solution by MHT concept for k solutions is $O(2^m k \log(2^m k))$ where m is the number of the body parts.

APPENDIX B

PUBLICATIONS (INTERNATIONAL CONFERENCE)

ROBUST CONTOUR TRACKING IN CLUTTERED BACKGROUND USING SNAKE ON WEIGHTED-GRADIENT IMAGE

Kittiya Khongkraphan[1,a], *Pakorn Kaewtrakulpong*[2,b]

Department of Computer Engineering[1]
Department of Instrumentation and Control System Engineering[2]
Faculty of Engineering, King Mongkut's University of Technology Thonburi
126 Prachautid, Bangmod, Toongkru, Bangkok, 10140, Thailand.
Email: kittiya_pn@hotmail.com[a], pakorn.kae@kmutt.ac.th[b]

ABSTRACT

In this paper, we focus on tracking human head in cluttered background. We use particle filtering with Snake algorithm. A so-called *weighted-gradient* image is introduced and supplied to Snake fitting. It gives more weights to moving object boundary obtained from a background subtraction technique and can reduce incorrect matching due to object texture and spurious edges. Experimental results showed robust tracking of human head in cluttered background. The approach is currently limited to a fixed camera platform; however, it can be applied to an active platform by replacing background subtraction with an optical flow method.

1. INTRODUCTION

Tracking has attracted much attention to researchers in computer vision. It is used to find locations of objects in image sequences. It is very useful in many applications, e.g. visual surveillance, facial animation, faces recognition and human-computer interaction. Various methods have been proposed for visual tracking in the past several years. Blob tracking [1] is a basic method directly derived from Radar tracking. It normally uses the result from background subtraction [1] as measurement to the tracker. Once the regions have been detected, they are supplied as detected objects to the tracker [1]. Optical flow is a tracking method which detects and tracks change of areas [2]. The areas are usually grouped into objects using motion segmentation. For feature tracking, it deals with detecting and matching features such as edges lines or corners [3]. Similar to the optical flow technique, the features are grouped into objects by motion segmentation. Model-based tracking uses objects models e.g. contour model [4-5] or appearance model [6] to find objects in images. The tracker is responsible for providing good initial seeds for the search.

In this paper, we focus on tracking human head in cluttered background. Accurate tracking of body parts is normally performed by contour tracking, since it can identify parts of object during the tracking process. Snake is widely used for object shape modeling and tracking. It is selected in our proposed method, as it does not require training process. However, contour tracking in cluttered background still has problem due to object texture and spurious edges.

1.1 Previous Work

There are several existing approaches to contour tracking. Deformable template approach [7-8] involves finding a model for the contours of the object to be tracked, and matching this representation in successive images from the video stream. Active contours were first introduced by Kass et al. in 1988, with a model called Snake [4]. It considers curves within image that can move under the influence forces within the curve itself and external force derived from image data. There are several papers which present Kalman trackers; ones of the most well-known are Blake et al. [9] and Brockett and Blake [10]. The key idea of the Kalman tracker is the same as that of the general Kalman filter. In particular, it requires a learned linear stochastic dynamical model which describes the evolution of the contour to be tracked. Assuming that the observation of the contour is corrupted by Gaussian noise, the conditional density of the contour given all past observations may be found, and then used to estimate the contour position. The condensation tracker, as outlined in Blake and Isard [11], is similar to the Kalman tracker, in that it assumes that a dynamical model describing contour motion is known, and that uncertain observations are made. However, it is more general: neither the dynamical system nor the observation process need be linear.

By employing multiple hypotheses together with a model of system dynamic of particle filtering, the method can track objects more reliably in cluttered background. Many researches apply the particle filter in tracking. Xu and Li [12] proposed a particle filter with both color and shape models for tracking human head with an ellipse model. Since the model does not accurately describe the contour of the object, to make the gradient estimate more useful in case of the inaccurate modeling, the gradient at pixel is established as the maximum gradient by a local search along a normal direction. Shen et al. [13] proposed a particle filter on an active contour. It can be used to track boundary of a complex cell. However, it still has problem due to spurious edges when applied to human head tracking in cluttered background. Figure 1 shows results from tracking in cluttered background. The results made incorrect tracking due to spurious gradient. Our approach introduces weighted-gradient map to solve the problem.

Figure 1 Sample results from sequence in cluttered background

1.2 Motivation

Recently, many visual tracking algorithms have been developed but reliable tracking in cluttered background is still a difficult problem. In this paper, we focus on the above-mentioned problem, i.e. existence of noise and spurious edges due to cluttered background. We use background subtraction for segmentation of moving objects and create a so-called *weight map*. Since correct object boundary is normally obtained from the moving object mask, the gradient along the mask should be given priority. The weight map is then applied to external energy computation of Snake algorithm. It gives more weights to moving object boundary and should be able to reduce incorrect matching due to object texture and spurious edges. Although, the mask of moving objects may not be correct due to imperfection of background subtraction, generally the weight map tends to give more weights to correct moving object boundaries.

The paper is organized as follows. After the general introduction, motivation and previous work are presented in this section. Our proposed method is described in Section 2. In Section 3, we present some experimental results. Finally, conclusion and future work are discussed in Section 4.

2. PROPOSED METHOD

The outline of our approach is shown in Figure 2. Firstly, the background subtraction is used to segment moving objects from static background. The binary foreground mask obtained from the method is then used to calculate the weight map. The weight map is applied to energy minimization of Snake algorithm. The particle filter provides initial seeds for Snake and the energy of Snake is used in the measurement process of particle filtering.

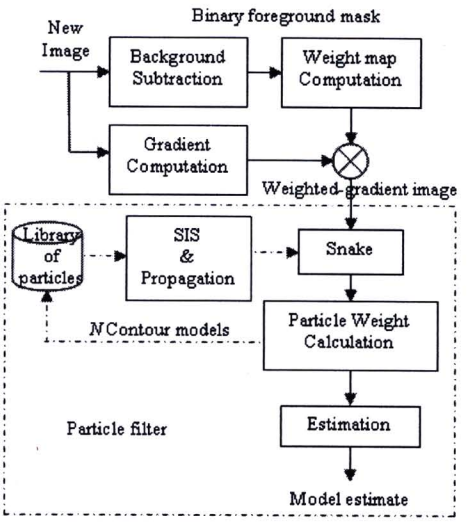


Figure 2 Flowchart of the contour tracking in cluttered background, (SIS – Sequence importance sampling)

2.1 Background Subtraction

Background subtraction is a process of subtracting the current image from the reference image to obtain moving objects. Our approach uses mixture of Gaussian distributions, in which the threshold is set to have more false positives than false negatives [1].

2.2 Weight Map Computation

The foreground mask from background subtraction is used in weight map computation. Boundary of object is extracted and Gaussian filter (of size 5 and σ^2 of 2) is applied. The weight map is obtained and is applied to energy of Snake.

2.3 Snake Algorithm

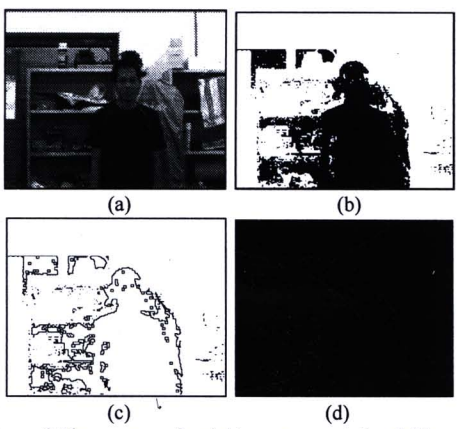
Snake is an edge-based model widely used for object shape modeling and tracking. It is simple and does not require training process. A general Snake is in a curve form $\boldsymbol{x}(s) = [\boldsymbol{x}(s), y(s)], s \in [0,1]$, which moves through the spatial domain of an image to minimize the energy function

$$E = \int_0^1 \frac{1}{2}(\alpha |\boldsymbol{x}'(s)|^2 + \beta |\boldsymbol{x}''(s)|^2) + E_{ext}(\boldsymbol{x}(s))ds \quad (1)$$

where α and β are weighting parameters that control the Snake's tension and rigidity, respectively. $\boldsymbol{x}'(s)$ and $\boldsymbol{x}''(s)$ denote first and second derivatives of $\boldsymbol{x}(s)$ with respect to s. The external energy function E_{ext} is derived from *weighted-gradient* image which is the multiplication of gradient image to the weight map,

$$E_{ext} = -|\nabla I(x,y)|^2 * W \quad (2)$$

where ∇ is the gradient operator and W is the weight map.



(a) (b) (c) (d)

Figure 3 The process of weight map computation (a) Image with moving object, (b) Binary foreground mask, (c) Boundary of objects, (d) Weight map

This gives more weights to pixels close to object boundary while external energy of background and object interior is reduced. By applying the method, incorrect matching due to object texture and spurious edges decreases significantly. The result from weight map computation is shown in Figure 3.

However, the initial contour should be close to the true boundary; otherwise, it will likely converge to wrong result. This initial contour is provided by the particle filtering.

Figure 4 (a) Gradient image (b) Weighted-gradient image

2.4 Particle Filtering

The particle filtering algorithm is introduced to solve non Gaussian noise and nonlinear contour tracking problem in image sequences [6]. It estimates current state from previous time steps and current measurement. In the approach to dynamic state estimation, the tracker attempts to construct posterior probability density function (*pdf*) of the state by a set of random samples with associated weights and to compute estimates based on these samples and weights. Since this *pdf* embodies all available statistics, it may be said to be the complete solution to the estimation problem. The probabilistic tracking composes of two phases: Prediction and Updating. The prediction step uses the system model to predict the state *pdf* ahead from one measurement time to the next. In updating state, the state from prediction step is updated by the latest measurement data.

The key idea of particle filtering is the approximation of the posterior density function $p(\mathbf{s}_t \mid \mathbf{z}_t)$ by a set of samples and correlation weights, $\{\mathbf{s}_t^{(i)}, w_t^{(i)} \mid i = 1,2,...,N\}$ where $\mathbf{s}_t^{(i)}$ is a sample of the state vector, the initialization of Snake, $w_t^{(i)}$ the corresponding weight, and $\mathbf{z}_t$ the observation vector. Each sample s represents one hypothetical state of Snake initialization with a corresponding discrete sampling probability w where $\sum_{i=1}^{N} w^i = 1$. The particle filter generates samples for the initialization of Snake algorithm based on prior result from previous images. The probability of each sample (or particle) is computed from energy of Snake. It is specified by a Gaussian distribution with variance σ^2,

$$w^{(i)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{Energy}{2\sigma^2})} . \quad (3)$$

where $w^{(i)}$ is the corresponding weight, σ standard deviation and *Energy* is energy from Snake. The position of the object is estimated at each time step by

$$E = \arg\max_{s^{(i)}} p(\mathbf{s}_t \mid \mathbf{z}_{1:t}) \approx \arg\max_{s^{(i)}, i=1,2,3,...,N} w^{(i)} \quad (4)$$

where E is the position estimate of the object, $\mathbf{s}_t^{(i)}$ the sample of the state vector and $w^{(i)}$ the corresponding weight.

3. EXPERIMENTS

To evaluate the performance of our proposed method, we tested our method to image sequences with simple and cluttered backgrounds. The experiments were conducted on 176x144 RGB image sequences. Figure 5 shows performance of tracking in a simple sequence, while Figure 6 shows tracking performance under relative cluttered background.

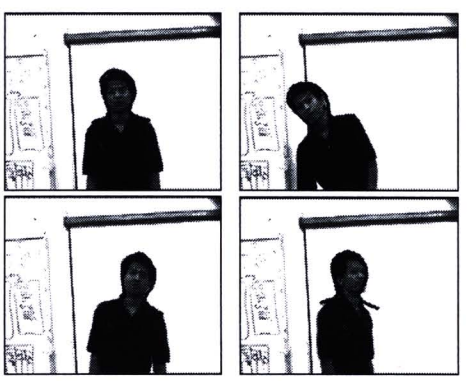


Figure 5 Sample results from simple sequence

Figure 6 Sample results from sequence with cluttered background

From experimental results, it can be seen that the application of weight map produces the best results. The approach has great influences on adjustment of the energy landscape of Snake in cluttered background. It gives more weights to moving object boundary so it can be able to reduce incorrect matching due to object texture and cluttered background.

The error of distance per control point is shown in Table 1. It shows comparison of errors in tracking using plain Snake and Snake with weighted-gradient image in simple sequence and cluttered background sequences.

Method	Simple Background	Cluttered Background
Snake	2.596	23.153
Snake on weighted-gradient image	0.896	1.595

Table1. Errors of distances (per control point) between Snake and Snake on weighted-gradient image

Nevertheless, problem of local minima still affects our method in a certain degree. An example of the problem is shown in Figure 7. To solve this problem, we introduce a lower limit for percentage of Snake model matching to our approach. This acts as a constraint to the energy minimization and the result can be seen in Figure 8.

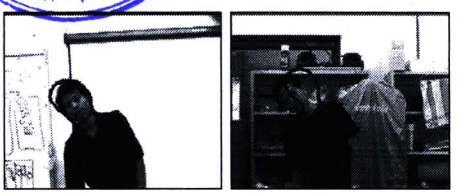


Figure 7 Local minima problem of moving Snake

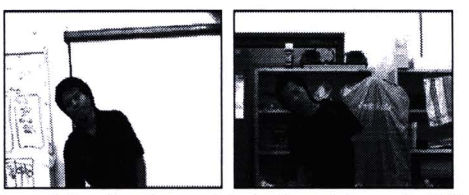


Figure 8 Results after an introduction of lower limit for percentage of Snake model matching constraint

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a contour tracking technique in cluttered background by applying weight map to energy minimization of Snake. In cluttered background, it gives more weights to pixels close to boundaries of moving objects. Although, the masks of moving objects may not be correct due to imperfection of background subtraction, generally the weight map tends to give more weights to correct moving object boundaries. This reduces incorrect matching due to object texture and spurious edges. This approach performs much better than previous contour tracking techniques.

However, the approach is currently limited to a fixed camera platform. For an active platform, our approach can be extended by replacing the background subtraction with an optical flow method. Future work also includes an investigation of the tracking in case of multiple objects.

5. REFERENCES

[1] P. Kaewtrakulpong, and R. Bowden, "A Real Time Adaptive Visual Surveillance System for Tracking Low-Resolution Colour Targets in Dynamically Changing Scenes", Image and Vision Computing, Vol. 21, No. 10, September, pp. 913-929, 2003.

[2] S. Basu, I. Essa , and A. Pentland , "Motion regularization for model based head tracking", Inter. Conf. on Pattern Recognition, Vienna, Austria, 1996

[3] Cox, I. J. and Hingorani S.L., "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking", IEEE Trans. PAMI, Vol. 18, No. 2, pp. 138-150, 1996

[4] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes:active contour models," International Journal of Computer Vision, pp. 321- 331, 1988.

[5] T: F. Cootes, C. J. Taylor, "Active Shape Models: *Smart Snakes*," in Proceeding British Machine Vision Conference, Springer-Verlag, 1992.
[6] T.F. Cootes, G.J. Edwards, C.J. Taylor, "Active Appearance Models," IEEE Trans. PAMI, Vol. 23, No. 6, pp. 681-684, 2001
[7] A. Yuille, P. Hallinan, and D. Cohen, "Feature extraction from faces using deformable templates," International Journals of Computer Vision, No.8., Vol. 2, pp. 99–112, 1992.

[8] P. Lipson and et al., "Deformable templates for feature extraction from medical images," In Proceeding 1st European Conference on Computer Vision, 1990.
[9] A. Blake, R. Curwen , and A. Zisserman, "A framework for spatio-temporal control in the tracking of visual contours," International Journals of Computer Vision, No. 11, Vol. 2, pp. 127–145, 1993.
[10] R. Brockett, and A Blake, "Estimating the shape of a moving contour," In Proc. of the 33rd IEEE Conference on Decision and Control, pp. 3247–3252, 1994.
[11] M. Isard, and A. Blake, "Condensation Conditional Density Propagation for Visual Tracking," International Journal of Computer Vision., 1996.
[12] Xu, X. and Li B., "Head Tracking using Particle Filter with Intensity Gradient and Color Histogram", IEEE International Conference, pp.888 – 891, 2005.
[13] A. Hailin Shen and et al, "Automatic tracking of biological cells and compartments using particle filters and active contours," Chemometries and Intelligent Laboratory System, Vol. 82, pp. 276 - 282, 2000

APPENDIX C

PUBLICATIONS (INTERNATIONAL JOURNAL)

IEICE Electronics Express, Vol. , No. , 1–6

A novel method of 2D articulated body tracking under self-occlusion and ambiguity

Kittiya Khongkraphan[1] and Pakorn Kaewtrakulpong[2,a]

[1] *Department of Computer Engineering*
[2] *Department of Control System and Instrumentation Engineering*
King Mongkut's University of Technology Thonburi
126 Prachautid, Bangmod, Toongkru, Bangkok ,Thailand ,10140

a) pakorn.kae@kmutt.ac.th

Abstract: A novel masker-less based method to track a 2D articulated body under self-occlusion and ambiguity in monocular image sequences is proposed. The proposed method applies SMC to both color and motion features. We employ a self-updated binary occlusion mask to increase accuracy in tracking. To alleviate the effect of illumination, each color body part is formed by a Gaussian mixture model in HS space, and the distribution intersection is used in distance measures of two probability distributions. Moreover, a motion cue is used to prune spurious solutions. Our technique can track the target reliably, especially in occlusion and ambiguous cases.

Keywords: Tracking, Occlusion, Ambiguity
Classification: Science and engineering for electronics

References

[1] J. Deutscher, A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering," Proc. CVPR, pp.126-133, 2000.
[2] W. Qu and D. Schonfeld, "Real-time decentralized articulated motion analysis and objecting from videos," IEEE trans. Image processing, pp. 2129-2138, 2007.
[3] A. Gritai, A. Basharat and M. Shah, "Geometric constraints on 2D action models for tracking human body," Proc. ICPR, 2008.
[4] D. Ramanan, D. Forsyth and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," Proc. CVPR, pp.271-278, 2005.
[5] E. B. Sudderth, M. I. Mandel, W. T. Freeman and A. S. Willsky, "Distributed occlusion reasoning for tracking with nonparametric belief propagation," Proc. NIPS, 2004.
[6] G. Hua, M. H. Yang and Y. Wu, "Learning to estimate human pose with data driven belief propagation," Proc. CVPR, pp. 747-754, 2005.
[7] K. Khongkraphan and P. Kaewtrakulpong, "Robust contour tracking in cluttered background using snake on weighted-gradient image," Proc. IWAIT, 2007.
[8] C. Ciaran, O. E. Noel and S. F. Alan, "Detector adaptation by maximizing agreement between independent data sources,"" Proc. IEEE In-

IEICE Electronics Express, Vol. , No. , 1–6

ternational Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2007.
[9] S. J. McKenna, Y. Raja and S. Gong, "Tracking color objects using adaptive mixture models," Image and vision computing, pp.225-231, 1999.
[10] A. Doucet, N. de Freitas, and N. Gordon, Sequential Monte Carlo Methods in Practice, Springer-Verlag, 2001.

1 Introduction

Tracking human body parts or poses has attracted the attention of many computer vision researchers. A number of algorithms have been proposed to address human body pose tracking, and one of the initial approaches was based on image-based approach. These approaches often fail under self-occlusion, where multiple human body parts occupy the same region of the input image. Moreover, several approaches generally suffer incorrect tracking due to ambiguity problems, since some symmetric body parts such as arms and legs normally have a similar appearance. Solving the self-occlusion and ambiguity problems is the focus of our paper.

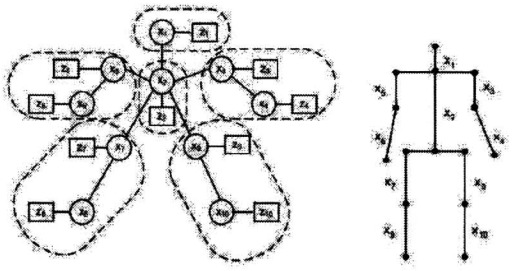
1.1 Previous work

To deal with the self-occlusion and ambiguity problems, a number of approaches utilize multiple cameras [1]. The main drawback of these approaches is the camera system setup. Several approaches focus on using strong prior knowledge of body poses; however, these are limited to normal poses in known activities such as walking [2] or aerobics [3]. Some approaches apply a binary occlusion mask to resolve the self-occlusion problem [4][5]. Such approaches require an assumption that the order of the limbs is known a priori. Some approaches use detected part candidates to help estimate postures approximately, even under self-occlusion [6]. Alternatively, some researches [2] form a human model using a collection of trees instead of a single tree [1]. The configuration of each tree can be obtained sequentially. Like the binary occlusion mask approach, these approaches require knowledge of limb order.

2 Proposed method

The main concept of our approach is to apply a binary occlusion mask and to decompose a human model into several subtrees to resolve the self-occlusion problem. Additionally, we combine both spatial and temporal information in the observation computation to deal with ambiguity problems. Firstly, an order of human body parts is determined for hierarchical tracking. Then the human body parts in the subtree closest to the camera are evaluated. The next step involves a binary occlusion mask being generated from that solution and then used in subsequent tracking of the other body parts according to their orders.

IEICE Electronics Express, Vol. , No. , 1–6



(a) (b)

Fig. 1. (a) The collection of subtrees of an articulated human body, (b)human body parts represented in the model.

2.1 Human body model

Similar to [2], we model a 2D view of a human body as a collection of M subtrees, which is denoted by $\mathbf{T} = \{T_k | k = 1, 2, ..., M\}$, where T_k is the k^{th} subtree. Each tree consists of hidden nodes and corresponding potentials, as shown by the circles and rectangles, respectively, in Fig. **1**(a). The human body parts represented by the nodes in the model are shown in Fig. **1**(b). The hidden nodes are represented by $X = \{\mathbf{x}_i | i = 1, 2, ..., N\}$, where $\mathbf{x}_i$ is the body part. Each part consists of five states: (x, y) position, orientation, width and length. A corresponding observation set is denoted by $Z = \{\mathbf{z}_i | i = 1, 2, ..., N\}$, where $\mathbf{z}_i$ is the image observation node for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by the observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$.

2.2 Human body order determination

We base our method on the assumption that orientation of the human face signifies the order of human body parts. To obtain the human facing, the human head is first detected [7] and then the facial skin is extracted from the head region [8]. A line separating the head region into two symmetrical parts is formed. The facing is then assigned to the side that has the maximum percentage of detected facial areas. In case of non-overlapping region, the facing is set to that obtained in the previous frame. A frame of a walking sequence and detected skin regions are displayed in Fig. **2**(a) and (b), respectively. The head contour and separating line are shown in Fig. **2**(c) in purple and black, respectively. The overlapping region between head and skin is shown in green in Fig. **2**(c).

3 Tracking

The head position is obtained from Section 2.2, whilst the remaining parts of the tree (torso, left arm, right arm, left leg and right leg) are individually tracked based on their orders. It is based on hierarchical tracking, and we apply a binary occlusion mask to alleviate detection of spurious features.

IEICE Electronics Express, Vol. , No. , 1–6

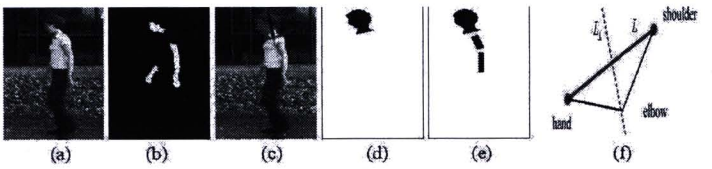


Fig. 2. (a) input image, (b) human skin, (c) human facing, (d) binary occlusion mask before left arm tracking, (e) binary occlusion mask after left arm tracking, (f) kinematics constraint.

Our approach uses the Sequential Monte Carlo (SMC) method [10] to track the human body. Each tree of the human model is considered a state space in SMC. The key idea of SMC is the approximation of the posterior density function $p(\mathbf{s}|\mathbf{z})$ by a set of samples (or particles) and correlation weight pair, $\{\mathbf{s}^i, w^i\}$; $i = 1, 2, ..., N$, where $\mathbf{s}^i$ is the i^{th} sample of the state vector, w^i its corresponding weight, $\mathbf{z}$ the observation vector and N the number of subtrees. The probability of each sample is computed from

$$w^i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(1-w^i_{c,m})^2}{2\sigma^2}}. \tag{1}$$

where w^i is the corresponding weight of the i^{th} sample, σ the standard deviation and $w^i_{c,m}$ the observation value obtained by combining color and motion distance measures. It can be computed from

$$w^i_{c,m} = \alpha w^i_c + (1 - \alpha) w^i_m. \tag{2}$$

where w^i_c, w^i_m are color appearance and motion scores, respectively. (The scores are normalized to be between 0 and 1.) α is the weight of blending between color and motion information. The sample with maximum weight is selected to be the solution

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}^i} p(\mathbf{x}|\mathbf{z}) \approx \arg\max_{\mathbf{s}_i} w^i; i = 1, 2, 3, ..., N. \tag{3}$$

where $\hat{\mathbf{s}}$ is the state estimate of the object, $\mathbf{s}^i$ the sample of the state vector and w^i the corresponding weight. After tracking, the binary occlusion mask is updated for tracking subsequent body parts. The binary occlusion masks of before and after tracking the left arm are shown in Fig. **2**(d) and (e), respectively.

3.1 Sample generation

The concept of sample generation is to create appropriate candidates for each joint position and evaluating them according to body-part model to obtain the best solution. The samples are generated based on body-part ratio of Leonardo Da Vinci's anatomy concept, prior knowledge of potential positions on thinning lines of detected body-part silhouette and of detected skin regions, and estimated body-part order.

IEICE Electronics Express, Vol. , No. , 1–6

Starting from neck and shoulder joints, an upside-down-T shape is fitted on the head-separated line (see Fig. **2**(c)) and samples are generated around its end and intersection points. Position of the hand closer to camera is estimated from potential end positions of the remaining thinning lines. For elbow, samples are limited by a kinematics constraint on the line $L_{\perp}$, as depicted in Fig. **2**(f). Its samples are generated around $L_{\perp}$. Another upside-down T-shape is fixed using the neck joint and the estimated torso location. Similarly, the samples of hip and abdomen are generated. Foot and knee positions are generated similarly to those of the arm.

4 Observation function

Generally, the observation function is used to measure the observation likelihood of the samples. A color feature is widely used in human tracking thank to their robustness to rotation in depth, shape changing and scale changing. However, it suffers color changing over time due to changes in scene illumination and visual angle. Another major problem is ambiguity due to the similarity of limb features. To alleviate the above problems, we integrate spatial and temporal information into the observation computation for color and motion features. For the color feature, we use hue (H) and saturation (S) components of the HSI space [9]. Lightness (I) is not used in the observation function so that it is robust against global illumination changes. Each human body is formed by a Gaussian mixture model $m(\mathbf{s})$ based on HS components.

$$m(\mathbf{s}) = \sum_{i=1}^{M} w^i \eta(\mathbf{s}; \mu^i, \Lambda^i). \tag{4}$$

where $\eta(\mathbf{s}; \mu^i, \Lambda^i)$ denotes the i^{th} Gaussian component with mean μ^i and covariance Λ^i. w^i is the i^{th} mixing weight satisfying $\sum_{i=1}^{M} w^i = 1$. From the color models of the sample $m(\mathbf{s})$ and the template $m(\mathbf{t})$, the similarity measure is defined by the distribution intersection between the color model of the sample and the template as

$$w_c = m(\mathbf{s}) \cap m(\mathbf{t}). \tag{5}$$

where w_c is the color score. (A discrete approximation of the intersection is implemented in our work.)

In motion feature, we also model the motion feature by a Gaussian to represent the distribution of the distance between predicted point and sample. The position of each body part is first predicted by a first-order motion and then used to compare with positions of samples.

$$w_m = \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{d^2}{2\omega^2}}. \tag{6}$$

where d is the distance between the positions of predicted and sampled positions, w_m the motion score and ω the standard deviation.

IEICE Electronics Express, Vol. , No. , 1–6

5 Experiments

To evaluate the performance of our proposed method, we tested it on two image sequences, namely, "walk straight" and "walk in a circle", containing 60 and 145 frames, respectively. Figs. 3(a) and (b) show the performance of the two sequences. The right limbs are shown in purple, whilst the left limbs are shown in green. The accuracy is evaluated from an averaged error distance compared with manually labeled positions, as 2.03 and 2.35 pixels/joint in "walk straight" and "walk in a circle", respectively.

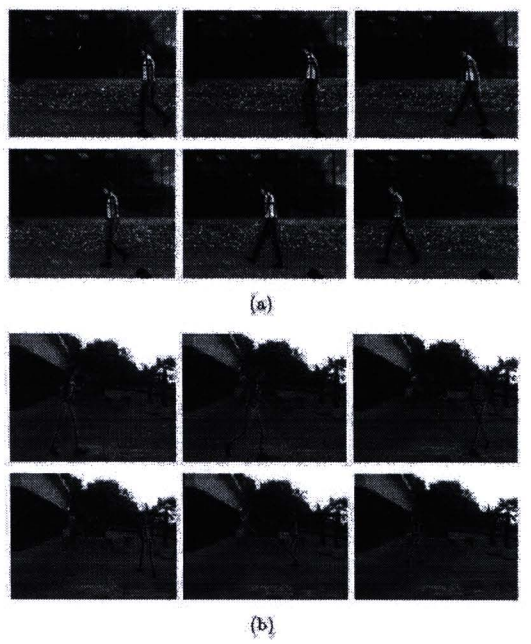


Fig. 3. Some results using our approach (a) "walk straight" (b) "walk in a circle".

6 Conclusion

We propose a masker-less based approach to track the human body. The proposed method applies SMC to both color and motion features. By employing a self-updated binary occlusion mask and a model of the observation computation, the method can track objects reliably in cases of self-occlusion and ambiguity. The measurement is based on a color cue modeled by the Gaussian mixture model in HS space, and the distribution intersection is used in distance measures of two probability distributions. To increase the performance of tracking in cases of ambiguity, a motion cue is used to prune spurious solutions. From our experiments, our proposed method performed very well, especially in occlusion and ambiguous cases.

IEICE TRANS. INF. & SYST., VOL.E94–D, NO.4 APRIL 2011

PAPER

Efficient Human Body Tracking by Quick Shift Belief Propagation

Kittiya KHONGKRAPHAN†, *Nonmember and* **Pakorn KAEWTRAKULPONG**†a), *Member*

SUMMARY We propose a novel and efficient approach for tracking 2D articulated human body parts. In our approach, the human body is modeled by a graphical model where each part is represented by a node and the relationship between a pair of adjacent parts is indicated by an edge in the graph. Various approaches have been proposed to solve such problems, but efficiency is still a vital problem. We present a new Quick Shift Belief Propagation (QSBP) based approach which benefits from Quick Shift, a simple and efficient mode seeking method, in a part based belief propagation model. The unique aspect of this model is its ability to efficiently discover modes of the underlying marginal probability distribution while preserving the accuracy. This gives QSBP a significant advantage over approaches like Belief Propagation (BP) and Mean Shift Belief Propagation (MSBP). Moreover, we demonstrate the use of QSBP with an action based model; this provides additional advantages of handling self-occlusion and further reducing the search space. We present qualitative and quantitative analysis of the proposed approach with encouraging results.

key words: *human body tracking, belief propagation, quick shift*

1. Introduction

Tracking human body parts or body pose has recently attracted increased attention from computer vision researchers. These approaches typically focus on the body parts in more detail than tracking human location as a whole. This is an important problem to solve because it can serve as a front end for higher-level processes to understand human activities in several applications, such as human-computer interaction, virtual reality, and character animation. A number of algorithms have been proposed to address human body pose tracking. One of the initial approaches was based on the use of *markers*; however, this method requires a great deal of user intervention. The system using markers is uncomfortable in many applications due to the use of special equipment which must be installed on the body. On the other hand, the marker-less system can run without special equipment.

Among the marker-less methods, several researchers turn to image-based approaches. In the image-based techniques, features from image (s) are served as its input to estimate human pose. There are two main approaches: discriminative and generative approaches. In the discriminative approach (or bottom up approach), candidate body parts are first extracted from the image and then used to generate possible configurations of human body. The performance of this approach is based on results of feature extraction since it is used to represent body parts.

In the generative approach (or top down approach), human is normally modeled as a skeleton model and the main concept is an effort to fit a predefined human model into image data. A large number of projected samples are generated and their similarity with respect to the input image is measured and compared. Such approaches are generally computationally intensive because of the complex search over the high dimension state space. The computational cost of these approaches is $O(N^m)$, where N is the number of samples for each body part and m is the number of body parts. Several approaches have been proposed to reduce the computation load. Constraints are normally introduced to limit the search space for some specific applications, e.g. walking parallel to a plane [1] or golf swing movements [2]. Motion models and prior knowledge are also employed to predict the state space trajectory or to reduce the search space [1]. Lee et al. [3] detected some body parts separately and used them to update subsets of the state space. Alternative approaches attempted to reduce the number of samples using support vector machines [4], annealed particle filtering [5], hybrid Monte Carlo filtering [6] and kernel particle filtering [7].

Several researchers turned to part-based approach that employs Bayesian inference concept to estimate human pose. It has been popular due to reduction in computational cost from $O(N^m)$ to $O(mN^2)$ [8]–[14]. In this approach, the human body is represented by a graphical model where each part is represented by a node and the relationship between the adjacent parts is indicated by an edge of the graph. Each body part is considered separately and then is combined into the global solution later. However, the computational time is still quite substantial, which limits its application from real-time human body tracking. Various approaches [15], [16] have been proposed to alleviate this problem, but efficiency is still a vital problem. The high-dimensional space is generally associated with exponential increase of computation time, and that is the motivation of our paper.

The main contribution of our paper is the introduction of a novel method for fast tracking of articulated body while maintaining high precision. The proposed QSBP model is based on the key idea of efficiently refining the estimate of the mode of marginal probability in each iteration of belief propagation. This leads to the reduction in compu-

Manuscript received May 13, 2010.
Manuscript revised November 8, 2010,

†The authors are with the Department of Control System and Instrumentation Engineering, King Mongkut University of Technology Thonburi 126 Prachautid, Bangmod, Toongkru, Bangkok, 10140, Thailand.
a) E-mail: pakorn.kae@kmutt.ac.th

DOI: 10.1587/transinf.E94.D.905

IEICE TRANS. INF. & SYST., VOL.E94–D, NO.4 APRIL 2011

tation time since it requires fewer iterations than MSBP (Mean Shift Belief Propagation) [16] and require fewer samples than NBP (Non-parameter belief propagation) [17]–[19], [26]–[29]. The samples only in the close proximity of the current estimate of mode are used. We utilize Quick Shift [20] for mode seeking method in each iteration to set initial samples. In addition, we also demonstrate the use of a model video, when possible; to further reduce the search space for QSBP (Quick Shift Belief Propagation). Using similar action based models have been shown to be useful for addressing self-occlusion problem [21] as a motion model.

The paper is organized as follows. After the general introduction, related work is presented in this section. Our proposed method is described in Sects. 2 and 3. In Sect. 4, we present some experimental results. Finally, the conclusion is discussed in Sect. 5.

1.1 Related Work

The part-based is an approach that is more powerful than others since it can reduce computational cost from $O(N^m)$ to $O(mN^2)$. This approach represents parts of the human body with a graphical model. The main idea is to pass messages iteratively between adjacent nodes of the graph. Two popular techniques used to calculate the messages are the belief propagation and mean field Monte Carlo methods [8]. Their difference lies in the pattern of messages. Shen at el. [9] compared and reported that belief propagation offered better performance than that of mean field Monte Carlo.

The tracking is performed by generating a number of candidate samples for each node and then calculating their beliefs. The beliefs are computed from observation functions and messages. Ramanan et al. [10], [11] proposed 2D human tracking by considering candidate parts first and then combining those to find the optimal solution by Belief Propagation (BP). Gao and Shi [12] detected human faces and hands by color cues to guide the sample generation for face and hand parts in order to achieve better efficiency in tracking. In some papers, both messages and beliefs (or marginal distributions) are represented by weighted samples [13], [14], whereas several other researchers represented messages with a more complex continuous distribution such as mixtures of Gaussians [17]–[19] or NBP. The samples in NBP are drawn by the Gibbs sampler for all iterations, which leads to a huge increase in computational time. To alleviate this problem of NBP, Han et al. [15] proposed to approximate the mixture of Gaussians by mode propagation and kernel filtering. They report that their method is 80 times faster than BP for tracking a 2D articulated body model; however, it is still far from adequate for real-time tracking requirements. Park et al. [16] proposed MSBP. They further reduce time complexity by computing only samples that moves toward the best solution. It is 30-50 times faster in 2D state vector tracking, and 300 times faster in 3D state vector tracking than BP.

To handle self-occlusion problem, some approaches utilize multiple cameras [5], [27]. The occlusion constraint is proposed in likelihood computation [28]. Wang and Mori proposed occlusion and spatial constraints by representation of human with multiple tree models [30]. Some approaches make use of the prior 2D and/or 3D information about the structure or kinematics of human body. Some shape based [22], motion-based [4], [23], and a combination of both approaches [24], [25] have been proposed in the literature. Their availability of large databases of shapes and motion patterns increases robustness to viewpoint change.

However, tracking 2D articulated human body parts is still a difficult problem with high computational cost. In this paper, we focus on this problem. We introduce a novel and efficient approach for tracking 2D articulated human body parts. It extends the belief propagation by applying mode seeking. As we will show in Sect. 2.3, the computational complexity of our proposed approach is approximately 36 and 99 times faster than those of MSBP and BP in case of 4-state (2D position plus body part length and angle) tracking. We also incorporate model video to limit our search space and to enhance tracking accuracy especially in case of occlusions.

1.2 Quick Shift

Quick Shift, proposed by Vedaldi and Soatto [20], is a simple and extremely efficient mode seeking method. Like Mean Shift, Quick Shift is a local optimization algorithm. Mean Shift can be regarded as a gradient ascent method [31] while the Quick Shift does not require gradient information. Quick Shift is a quick Euclidean version of medoid shift that is guaranteed to converge for all starting locations [32].

In mode seeking techniques, it is started by defining the multivariate kernel density estimate. Like these techniques, Quick Shift also starts by computing the kernel density estimate

$$f(\mathbf{a}) = \frac{1}{M}\sum_{i=1}^{M}\varphi(\mathbf{a}-\mathbf{a}_i), \tag{1}$$

where $\mathbf{a}_i$ is the i^{th} data point and $\mathbf{a}_i \in X = \mathbb{R}^d$, $\varphi(\mathbf{a})$ is a kernel function [32] (e.g. Gaussian) and M is the number of data points. The main concept of Quick Shift is the iterative movement of each mode estimate from its current position to a new position which is the nearest neighbor with higher probability until a mode is reached. The updated position of data point $\mathbf{a}_i$ at iteration $k+1$ is computed by

$$\mathbf{y}_i^{k+1} = \underset{\mathbf{a}_j \in \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M\}: P(\mathbf{a}_j) > P(\mathbf{y}_i^k)}{\arg\min} D(\mathbf{y}_i^k, \mathbf{a}_j),$$

$$P(\mathbf{b}) = \frac{1}{M}\sum_{j=1}^{M}\varphi(D(\mathbf{b}, \mathbf{a}_j)), \tag{2}$$

where $D(\mathbf{y}_i, \mathbf{a}_j)$ is the distance between current positions of y_i^k and data point $\mathbf{a}_j$, $P(\mathbf{a}_i)$ is probability value of data point $\mathbf{a}_i$. The mode seeking is repeated on $\{\mathbf{y}_i^k\}$ until no further change in labelling occurs and then the modes are obtained

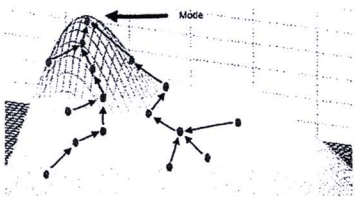


Fig. 1 Probability surface and motion of data points toward mode value of Quick Shift. Motion directions and data points are plotted by arrows and dots, respectively.

as the set of unique locations

$$mode = \{\mathbf{y}_i^l\}, \tag{3}$$

where $\mathbf{y}_i^l$ is the final position of $\mathbf{y}_i$.

All data points (initial values of mode estimates) move toward a single mode as shown in Fig. 1. To balance under- and over-fragmentation of the modes, a threshold parameter, κ, is introduced into Eq. (2)

$$y_i^{k+1} = \underset{\mathbf{a}_j \in \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M\}: P_j(\mathbf{a}_j) > P_i(y_i^k), D(y_i^k, \mathbf{a}_j) < \kappa}{\arg\min} D(y_i^k, \mathbf{a}_j). \tag{4}$$

2. Proposed Method

At each frame, a silhouette of the human body is obtained by a background subtraction. The silhouette as well as its original image are served as the input in our tracking method. The main concept of our approach is applying mode seeking to reduce computational time. A model video concept [21] is used for initializing the motion model in our proposed method. The initial state is served as an initial configuration for our QSBP. For the next step, samples are generated around the initial state and their probabilities are measured by belief propagation. Only the best solution of each node is selected to be a part of the optimal pose solution. We applied Quick Shift with belief propagation for mode seeking in each body part so that all samples (initial values of the mode estimates) are not necessary moved to a single point. In this sense, it reduces risks of getting into spurious solutions which have highest probability. From our approach, the modes are selected as initial samples in the next iteration of belief propagation.

2.1 Human Model

We model a 2D view of a human body as a graphical model with N hidden nodes and pair-wise potentials as shown in Fig. 2 (a). The hidden nodes are represented by $\mathbf{X} = \{\mathbf{x}_i | i \in [1, N]\}$, where $\mathbf{x}_i$ is the i^{th} body part consisting of 4 states, i.e., position coordinate, orientation and length. A corresponding observation set is denoted by $\mathbf{Z} = \{\mathbf{z}_i | i \in [1, N]\}$, where $\mathbf{z}_i$ is the image observation node for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by an observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$. In addition, every pair of adjacent body parts $\mathbf{x}_i$ and $\mathbf{x}_j$ (as defined by the structure), is connected and encoded by a potential function, $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$. Human body parts represented in the model are shown in Fig. 2 (b).

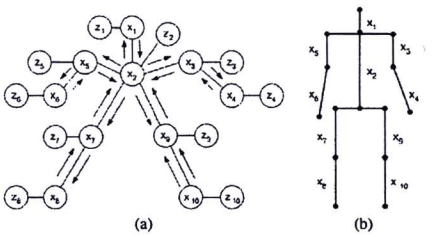


Fig. 2 The skeleton human model in our approach: (a) hidden nodes and pair-wise relationship and (b) joints and segments.

2.2 Model Video

A model video [21] is utilized for initializing the motion model in our approach. The main concept of the model video is to estimate the joint locations in the test video automatically using two geometric constraints. Given the locations of joints in the model video and the first frame of test video, the affine constraint is used to estimate initial positions based on invariance ratio between model video and test video. Moreover, the epipolar constraint is used to reduce estimation error of different actors and view points in the model and test videos.

One advantage of this work is the avoidance of error propagation from frame to frame in the estimation process, because each joint estimate is computed based on correspondence between the first frame of the model and the test videos. Another advantage is robustness to variations in anthropometry, execution rate, viewpoint and execution style. Moreover, it is not computationally expensive and does not require extensive training. We have found from empirical studies that a good tracking performance can be obtained if the model video gives an initial state within 1.5 times of the grid size. Figure 3 (a) shows the input frame. The candidate joints from the model video [21] are generated and overlaid on the input image as displayed in Fig. 3 (b). The best match with the silhouette is selected to be the initial samples as can be seen in Fig. 3 (c).

2.3 Quick Shift Belief Propagation

The belief propagation algorithm is an iterative method to infer the hidden state until it converges to the optimal solution. The marginal probability of $\mathbf{x}_i$ at iteration n, $p^n(\mathbf{x}_i|\mathbf{Z})$ can be computed by taking the product of incoming messages and local observations, as shown in Eq. (5). The incoming messages also contain prior knowledge of the node obtained from the neighboring nodes, as shown in Eq. (6).

$$p^n(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha \phi_i(\mathbf{x}_i, \mathbf{z}_i) \prod_{j \in \Gamma(i)} m_{ji}^n(\mathbf{x}_i). \tag{5}$$

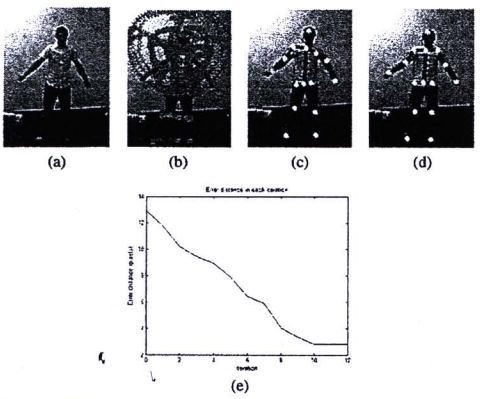


Fig. 3 QSBP tracking (a) input image, (b) predicted points from model video, (c) initial configuration, (d) final tracking result, (e) error distance in each iteration.

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i^{th} hidden and corresponding observation nodes, respectively. $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$ is the observation function of node i. α is a normalizing factor. For graphical models with continuous hidden state, $m_{ji}^n(\mathbf{x}_i)$ is a message sent from node j to i at iteration n and can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \int_{x_j} (\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)\phi_j(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j))d\mathbf{x}_j. \quad (6)$$

where $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$ is the potential function between nodes i and j, $\phi_j(\mathbf{x}_j, \mathbf{z}_j)$ is the observation function of node j and $\Gamma(j)\backslash i$ represents all the neighboring nodes of j except node i.

In our approach, we define parts of human body by nodes of graph that the optimal hidden node is computed by maximizing the marginal probability of each node given the current observation. Instead of considering all the possible states of our nodes, we work on a grid around initial samples (modes found in the previous iteration). A new discrete grid is generated around the mode. The grid is 5x5x3x3 of 4 states (x, y positions, length and angle of body part). The marginal probability of $\mathbf{x}_i$, $p(\mathbf{x}_i|\mathbf{Z})$ can be computed by taking the product of incoming messages and local observations, as shown in Eq. (5). The incoming messages also contain prior knowledge of the node obtained from the neighboring nodes. The message sent from node j to node i at iteration n, $m_{ji}^n(\mathbf{x}_i)$, can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \sum_{x_j} \left(\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)\phi_j(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j) \right). \quad (7)$$

The message at the first iteration, $m_{ji}^0(\mathbf{x}_i)$, is defined to be 1. From the iterative concept of belief propagation, the best solution is obtained in the final iteration by

$$\mathbf{x} = \arg\max_{s_i\in\{s_1,s_2,\ldots,s_N\}} p^n(\mathbf{s}_i|\mathbf{Z}), \quad (8)$$

where N is the number of samples in the grid. Figure 3 (d) shows the best solution in the final iteration of tracking and Fig. 3 (e) illustrates error distance in each iteration of tracking using our approach. It can be seen that the error distance is reduced until reaching the best solution.

2.4 Mode Seeking in Belief Propagation

In mode seeking, we use marginal probability of belief propagation in movement of mode estimates. Only samples around the modes are computed, not the entire surface of belief propagation. This makes the convergence to an optimal solution very fast and computation time is reduced. The updated position of sample $\mathbf{s}_i$ at iteration $k+1$ of the mode seeking is computed by

$$\mathbf{y}_i^{k+1} = \arg\min_{\mathbf{s}_j\in\{\mathbf{S}_1,\mathbf{S}_2,\ldots,\mathbf{S}_N\}:P(\mathbf{S}_j|\mathbf{Z})>P(\mathbf{y}_i^k|\mathbf{Z}),D(\mathbf{y}_i^k,\mathbf{s}_j)<\kappa} D(\mathbf{y}_i^k, \mathbf{s}_j), \quad (9)$$

where $D(\mathbf{y}_i^k, \mathbf{x}_j)$ is the distance between current positions of $\mathbf{y}_i^k$ and sample $\mathbf{x}_j$ which is less than a distance threshold κ and $P(\mathbf{x}_i|\mathbf{Z})$ is the marginal probability value of sample i. From Quick Shift mode seeking concept, the position of $\mathbf{y}_i$ is updated until no further change in labelling occurs and then modes of samples are obtained as a unique set

$$mode = \{\mathbf{y}_i^l\}, \quad (10)$$

where y_i^l is the final position of $\mathbf{y}_i$.

We find the mode of marginal probability by our approach which is faster than by MSBP [16]. Because moving toward a mode of Quick Shift is based on the locally maximum probability value, while the Mean Shift is based on weighted mean value. In this sense, the number of iterations in moving of Quick Shift approach is less than that of Mean Shift which makes its converge to the optimal solution much faster than Mean Shift. Figures 4 (a) and (b) show convergence to the optimal solution by QSBP and MSBP, respectively. From these figures, each initial sample is plotted by a solid square marker. The only grid members (plotted by markers inside of a grid window) around the initial sample are considered in moving toward a mode of the sample. The new position of the mode estimate in each iteration until they reach a mode are shown by a circle marker.

Main steps of proposed approach

1. Generate initial joint predictions from model video [21] as shown in Fig. 3 (b).
2. Select the best match to be initial samples as shown in Fig. 3 (c).
3. Iterate steps 4-7 until convergence.

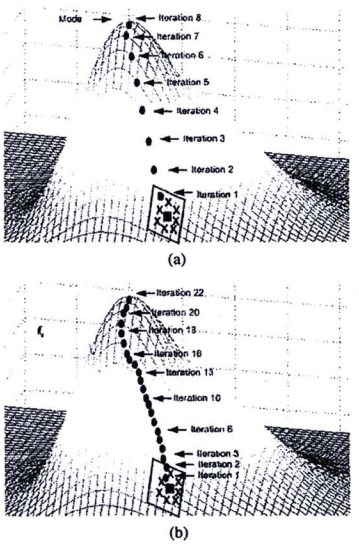


Fig. 4 A comparison of moving toward a mode value by (a) Quick Shift, and (b) Mean Shift.

4. Generate a local grid of samples around each initial sample.
5. Compute message by Eq. (7).

$$m_{ji}^{n}(\mathbf{x}_i) \leftarrow \sum_{x_j}\left(\psi_{ji}(\mathbf{x}_i,\mathbf{x}_j)\phi_j(\mathbf{x}_j,\mathbf{z}_j)\prod_{k\in\Gamma(j)\setminus i} m_{kj}^{n-1}(\mathbf{x}_j)\right).$$

6. Compute marginal probability using in Eq. (5).

$$p^{n}(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha\phi_i(\mathbf{x}_i,\mathbf{z}_i)\prod_{j\in\Gamma(i)} m_{ji}^{n}(\mathbf{x}_i).$$

7. Seek mode using Quick Shift, set each mode as initial sample for the next iteration.
8. Select the best solution as

$$\mathbf{x} = \underset{x_i\in\{x_1,x_2,\ldots,x_N\}}{\arg\max}\ p^{n}(\mathbf{x}_i|\mathbf{Z}),$$

2.5 Observation Function and Potential Function

The observation function $\phi_i(\mathbf{x}_i,\mathbf{z}_i)$ is used to measure the joint likelihood of $\mathbf{z}_i$ and $\mathbf{x}_i$. In order to measure the likelihood, each body part $\mathbf{x}_i$ is modelled by a planar patch and projected onto the input image, and then its likelihood (or similarity) is computed. In this work, region overlapping [3], [5] and RGB color are the features used for similarity measurement. The region overlapping feature of node $\mathbf{x}_i$ is computed by

$$w_i = \frac{1}{2}\left(\frac{n_{i,o}}{n_{i,p}} + \frac{n_{i,o}}{n_m}\right), \quad n_m = \underset{i}{\arg\max}\ n_{i,o} \tag{11}$$

where $n_{i,o}$ is the number of pixels in overlapping region between projected region of i^{th} sample and the silhouette. $n_{i,p}$ is the number of pixels in the i^{th} projected region. This measure is very simple and efficient; however, its reliability reduces greatly in case of occlusion. This is because overlapped parts form a larger foreground region which provides high values of this feature in several false locations.

In case of occlusion, we therefore switch to using a more detailed color feature (Sect. 3 explains how to detect the beginning and the end of occlusion). For the color feature, each human body part is formed by a color histogram based on RGB components. From the color models of the node $c(\mathbf{x}_i)$ and the template, t, the similarity measure is defined by the histogram intersection between the color shape matching model of the sample and the template as

$$w_i = \sum_{j=1}^{n} c_{j,i} \cap t_j, \tag{12}$$

where $c_{j,i}$ is the normalized number of pixels in the j^{th} bin of the i^{th} node and t_j is the normalized number of pixels in the j^{th} bin of template in RGB color histogram and n is the number of bins in each histogram. The observation function $\phi_i(\mathbf{x}_i,\mathbf{z}_i)$ of node $\mathbf{x}_i$ is computed by

$$\phi_i(\mathbf{x}_i,\mathbf{z}_i) = \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{(1-w_i)^2}{2v^2}}, \tag{13}$$

where w_i is the similarity of node $\mathbf{x}_i$ obtained by Eq. (11) or Eq. (12) and v is its standard deviation. For a low value of v, a more weight is given to the appearance similarity.

The $\psi_{ji}(\mathbf{x}_i,\mathbf{x}_j)$ potential function is used to show the relationship between body parts i and j. We model the potential function by a Gaussian to represent the distribution of the Euclidean distance between the two adjacent body parts.

$$\psi_{ji}(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{D(\mathbf{x}_i,\mathbf{x}_j)^2}{2\sigma^2}}, \tag{14}$$

where $\mathbf{x}_i$ is the i^{th} body part, $\psi_{ji}(\mathbf{x}_i,\mathbf{x}_j)$ is a potential function of body parts i and j, $D(\mathbf{x}_i,\mathbf{x}_j)$ is the Euclidean distance between the connected points of body parts i and j, and σ is its standard deviation. In the same way as v, σ specifies insensitivity to displacement of adjacent body parts.

3. Occlusion Handling

A common problem in human body tracking is self-occlusion problem. To reduce the computational complexity and handle self-occlusion problem of tracking in our method, we used prediction information from a model video [21] as the motion model. It is used for sample generation in the first iteration. We use two measures for detecting start and end of self-occlusion as in [21]. The first measurement is α_j^t, which represents the area of the foreground silhouette, corresponding to the j^{th} segment in the t^{th} frame.

The second measure is β_j^t, which represents the proportion of the detected segment j that is occluded by the other segments of the cardboard model. The condition for occlusion is based on the normalized change over time τ,

$$\frac{\sum_{i=t-\tau}^{t-1} \alpha_j^{i+1} - \alpha_j^i}{\tau \alpha_j^{t-\tau}} < T \ , \quad \frac{\sum_{i=t-\tau}^{t-1} \beta_j^{i+1} - \beta_j^i}{\tau \beta_j^{t-\tau}} > T. \qquad (15)$$

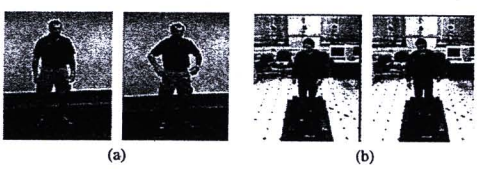
where T is the percentage threshold and τ is the number of previous frames that are used to consider occlusion. Positive T value signifies occlusion entering, while the negative T value indicates occlusion termination.

4. Experiments

To evaluate the performance of our proposed method, we tested it on several videos and compared it with the BP and MSBP methods in both simple and occlusion cases. We tested our method on 6 sequences of UCF dataset, containing 400, 162, 400, 560, 500 and 500 frames. These videos included aerobics style activities that were also used in [21]. Moreover, we experimented on 2 sequences of CMU dataset (also used in [16]), containing 199 and 190 frames including walking action in both front and side views. These videos include both simple and self-occlusion cases. To present qualitative and quantitative results, we compared our approach with [16]. Note that the joint locations were manually initialized in the first frame, and then an RGB template of each human part is automatically generated from the initialized joint locations. The sample prediction was then performed automatically for the remaining frames like the method presented in [21].

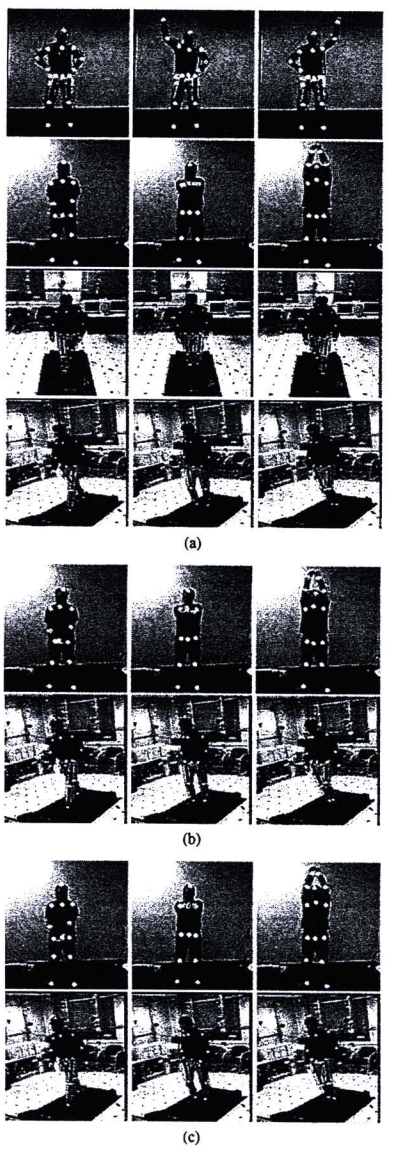
In our experiments, we generated samples with a size of 8,000 samples for BP, and a 5×5×3×3 grid for both MSBP and QSBP, respectively. Those methods were run until convergence (joint position movement is less than 2 pixels or 50 iterations are reached.) The threshold parameter in mode seeking of Quick Shift and the kernel size of MSBP were chosen as half of the grid size. For occlusion handling, we used 70 percentage in our experiments and the number of pervious frames was chosen as 15 to consider occlusion. The bin size of RGB histogram was 16×16×16. In observation function, the standard derivation of observation function and potential function were chosen as 0.4 and 3, respectively.

We applied model video to the sample prediction in the first iteration of belief propagation. The samples of model video on UCF and CMU datasets are shown in Figs. 5 (a) and (b), respectively. Samples of tracking results are displayed in Fig. 6. The figure shows samples of human model fitting on 4 sequences. The fitting results of QSBP, BP and MSBP approaches are shown in Figs. 6 (a), (b) and (c), re-



(a) (b)

Fig. 5 The samples of model video (a) UCF dataset (b) CMU dataset.



(a) (b) (c)

Fig. 6 Some results of human body tracking by using model video in sample prediction. Similar results from (a) Quick Shift belief propagation, (b) Belief propagation and (c) Mean Shift propagation.

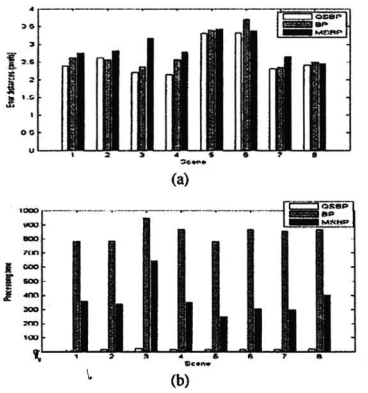


Fig. 7 A performance comparison between the proposed method and BP and MSBP are shown by white, gray and black bars, respectively (a) Accuracy (b) Efficiency.

spectively. Since similar results among all methods were obtained, only two results were included for BP and MSBP in Figs. 6 (b) and (c). It can be seen that the model fits well to the images for each approach. The results show that the accuracy of QSBP is comparable to those of BP and MSBP as illustrated in Fig. 7 (a). However, the computational time of QSBP is significantly less than those of BP or MSBP (Fig. 7 (b)). In particular, for our case of 4-state (2D position plus body part length and angle) tracking, our proposed technique is respectively 36 and 99 times faster than those of MSBP and BP. On average, the numbers of iterations to get the best solution are 9, 30 and 42 for QSBP, MSBP and BP, respectively. The 2D tracking results are compared with ground truth which is manually obtained. The average distance errors from the corresponding ground truth are shown in Fig. 7 (a). It can be seen that all methods provide accuracy; however, our approach is far more efficiency than the others as shown in Fig. 7 (b).

5. Conclusion

We propose a part-based tracking algorithm by integrating Quick Shift, a simple and efficient mode seeking method, into the belief propagation framework. The main idea is to find the mode of marginal probability of belief propagation to be used to predict points in the next iteration, and that way only samples around modes are computed in each iteration of belief propagation. Therefore, our proposed method needs fewer samples than NBP or MSBP. In addition, it converges to the best solution faster than the other methods. This approach can reduce the computational complexity dramatically due to the reduction of search space while preserving accuracy. In addition, we apply model video in the first iteration of belief propagation for predicting and resolving occlusion problems. The method was experimented on several videos and the results showed very good performance and robustness in both accuracy and efficiency.

Acknowledgement

This work is supported by Office of the Higher Education Commission (OHEC) and the Nation University Research (NRU) program. This work was conducted at UCF Computer Vision. The authors want to thank Professor Mubarak Shah and Arslan Basharat for fruitful discussion on various ideas in this paper.

References

[1] Z. Zhou, A. Prugel-Bennett, and R.I. Damper, "A Bayesian framework for extracting human gait using strong prior knowledge," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.11, pp.1738–1752, Nov. 2006.

[2] R. Urtasun, D.J. Fleet, and P. Fua, "Monocular 3D tracking of the golf swing," Proc. CVPR, pp.932–938, June 2005.

[3] M.W. Lee, I. Cohen, and S.K. Jung, "Partile filter with analytical inference for human body tracking," Proc. Workshop on Motion and Video Computing, pp.159–165, Dec. 2002.

[4] H. Sidenbladh, "Detecting human motion with support vector machines," Proc. ICPR, pp.188–191, Aug. 2004.

[5] J.D.A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering," Proc. CVPR, pp.126–133, 2000.

[6] K. Choo and D.J. Fleet, "People tracking using hybrid Monte Carlo filtering," Proc. CVPR, pp.321–328, 2001.

[7] C. Chang and R. Ansari, "Kernel particle filter: Iterative sampling for efficient visual tracking," Proc. ICIP, pp.977–980, Setp. 2003.

[8] Y. Wu, G. Hua, and T. Yu, "Tracking articulated body by dynamic Markov network," Proc. ICCV, pp.1094–1101, Oct. 2003.

[9] C. Shen, A.V. Hengel, A. Dich, and M.J. Brooks, "2D articulated tracking with dynamic Bayesian networks," Proc. Computer and Information Technology., pp.130–136, Sept. 2004.

[10] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.1, pp.65–81, Jan. 2007.

[11] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," Proc. CVPR, pp.271–278, June 2005.

[12] J. Gao and J. Shi, "Multiple frame motion inference using belief propagation," Proc. Automatic Face and Gesture Recognition, pp.875–880, May 2004.

[13] G. Hua and Y. Wu, "Multi-scale visual tracking by sequence belief propagation," Proc. CVPR, pp.826–833, July 2004.

[14] G. Hua, M. Hsuan, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," Proc. CVPR, pp.747–754, June 2005.

[15] T.X. Han, H. Ning, and T.S. Huang, "Efficient nonparametric belief propagation with application to articulated body tracking," Proc. CVPR, pp.214–221, June 2006.

[16] M. Park, Y. Liu, and R.T. Collins, "Efficient mean shift belief propagation for vision tracking," Proc. CVPR, pp.1–8, June 2008.

[17] M. Isard, "Pampas: Real-valued graphical models for computer vision," Proc. CVPR, pp.613–620, June 2003.

[18] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willky, "Nonparametric belief propagation," Proc. CVPR, pp.605–612, 2003.

[19] E.B. Sudderth, M.I. Mandel, W.T. Freeman, and A.S. Willsky, "Visual hand tracking using nonparametric belief propagation," Proc. CVPRW, pp.189–196, June 2004.

[20] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," Proc. ECCV, pp.705–718, Aug. 2008.

[21] A. Gritai, A. Basharat, and M. Shah, "Geometric constraints on 2D action models for tracking human body," Proc. ICPR, pp.1–4, 2008.

[22] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial structures for object recognition," Int. J. Comput. Vis., vol.61, no.1, pp.55–79, 2005.
[23] G.R. Bradski and J.W. Davis, "Motion segmentation and pose recognition with motion history gradients," J. Machine Vision and Application, pp.174–184, 2002.
[24] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell., vol.3, no.23, pp.257–267, March 2001.
[25] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detection," Proc. ICCV, pp.90–97, 2005.
[26] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," Proc. CVPR, pp.421–428, 2004.
[27] L. Sigal, M. Isard, Sigelman, and M. Black, "Attractive people: Assembling losse-limbed models using non-parametric belief propagation," Proc. NIPS, 2003.
[28] L. Sigal and M. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," Proc. CVPR, pp.2041–2048, 2006.
[29] E. Sudderth, M.I. Mandel, W.T. Freeman, and A.S. Willsky, "Distribution occlusion reasoning for tracking with nonparametric belief propagation," Proc. NIPS, 2004.
[30] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," Proc. ECCV, pp.710–724, 2007.
[31] C. Yizong, "Mean shift, mode seeking, and clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.8, pp.790–799, Aug. 1995.
[32] Y.A. Sheikh, E.A. Khan, and T. Kanade, "Mode-seeking by medoidshifts," Proc. ICCV, pp.1–8, 2007.

Kittiya Khongkraphan received the B.Sc. degree in Mathematics and M.Sc. degree in Computer Science from Prince of Songkla University, Thailand, in 1991 and 2000, respectively. Currently, she is a Ph.D. candidate of King Mongkut's University of Technology Thonburi, Thailand. Her research interests are computer vision and image processing.

Pakorn Kaewtrakulpong received the B.Eng. degree in Electrical Engineering from King Mongkut's University of Technology Thonburi (KMUTT) in 1992. He received M.Sc. and Ph.D. degrees in Systems Engineering from Brunel University in 1998 and 2002, respectively. He is currently an associate professor at the faculty of engineering, KMUTT. His research interests include machine vision, industrial automation and instrumentations.

IEICE TRANS. INF. & SYST., VOL.Exx–D, NO.xx XXXX 200x

PAPER

A novel reconstruction and tracking of 3D-articulated human body from 2D point correspondences of a monocular image sequence

Kittiya KHONGKRAPHAN[†], *Nonmember and* **Pakorn KAEWTRAKULPONG**[†a)], *Member*

SUMMARY A novel method is proposed to estimate the 3D relative positions of an articulated body from point correspondences in an uncalibrated monocular image. It is based on a camera perspective model. Unlike previous approaches, our proposed method does not require camera parameters or a manual specification of the 3D pose at the first frame, nor does it require the assumption that at least one predefined segment in every frame is parallel to the image plane. Our work assumes a simpler assumption, for example, the actor stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane in the first frame. Input into our algorithm consists of a topological skeleton model and 2D position data on the joints of a human actor. By geometric constraint of body parts in the skeleton model, 3D relative coordinates of the model are obtained. This reconstruction from 2D to 3D is an ill-posed problem due to non-uniqueness of solutions. Therefore, we introduced a technique based on the concept of multiple hypothesis tracking (MHT) with a motion-smoothness function between consecutive frames to automatically find the optimal solution for this ill-posed problem. Since reconstruction configurations are obtained from our closed-form equation, our technique is very efficient. Very accurate results were attained for both synthesized and real-world image sequences. We also compared our technique with both scaled-orthographic and existing perspective approaches. Our proposed method outperformed other approaches, especially in scenes with strong perspective effects and difficult poses.

key words: *perspective model, human skeleton, 3D tracking, MHT.*

1. Introduction

The image reconstruction of 3D-articulated humans has recently attracted increased attention from computer vision researchers for applications such as human-computer interaction, virtual reality and character animation. A number of methods have been proposed for estimating the 3D positions of a human body. The simplest way is by using special equipment such as 3D scanners; however, these systems are not only costly but also inconvenient for many applications. Because of their limitations, a number of researchers have turned to image-based approaches. In such approaches, the image serves as input to estimate 3D human positioning. According to the number of cameras, these models can be divided into two main groups: multi-view and single view. In the multi-view group, images captured from multiple calibrated cameras are used to reconstruct a 3D human pose by minimizing a score computed from fitting different pose configurations onto a 3D human body model [1], [2]. The main drawback of this approach relates to camera system setup. An alternative approach uses a monocular image to reconstruct a 3D-articulated body.

In the single view group, a monocular image is used in the reconstruction of a 3D human pose based on a camera-model concept. This approach attempts to lift 2D data to construct a 3D model of a human pose, normally formed as a skeleton model. The 2D joint correspondence positions that serve as input are usually obtained by hand labeling [3], [5], [7], matching to a marked image [8], or detecting from installed markers [4]. Assumptions or prior knowledge about the human pose or the imaging environment are normally made in most research studies. For example, various assumptions were employed: that most human parts are close to a plane parallel to the image plane [3], [7], [10], that at least one predefined human part is parallel to the image plane [12], [16], that root joint of human body moves parallel to the image plane without Z direction displacement [11] or that camera parameters are known a priori [5], [15], [16].

Based on the perspective concept, this paper proposes a novel method for estimating and tracking up-to-scale 3D positions of an articulated body from point correspondences in uncalibrated image sequences. We address the reconstruction of the 3D body in cases of strong perspective effects and arbitrarily complicated poses. We assume a simpler pose, for example, that a human stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane in the first frame. Because we provide a closed-form solution for 3D reconstruction, our method is very efficient. This work also alleviates the problems associated with the ill-posed reconstruction problem by applying a technique based on the concept of multiple

Manuscript received June 14, 2010.
Manuscript revised October 8, 2010.
Final manuscript received September 9, 9999.

[†]The authors are with Department of Control System and Instrumentation Engineering, King Mongkuts University of Technology Thonburi 126 Prachautid, Bangmod, Toongkru, Bangkok ,Thailand ,10140
a) E-mail: pakorn.kae@kmutt.ac.th

DOI: 10.1587/transinf.E0.D.1

hypothesis tracking (MHT) in selecting the solution. It is very efficient and can recover from the problem of selecting false solutions in arbitrarily complicated poses.

This paper is organized as follows. Following this initial introduction, Section 1 of this paper continues with a presentation of previous approaches of 3D construction using camera-based methods. Sections 2 and 3 describe our proposed 3D reconstruction and 3D body tracking methods, respectively. In Section 4, we present some experimental results. Finally, conclusions are drawn in Section 5.

1.1 Previous 3D reconstruction methods

There are two state-of-the-art approaches in the camera-model-based group, namely, scaled-orthographic and perspective approaches. The scaled-orthographic

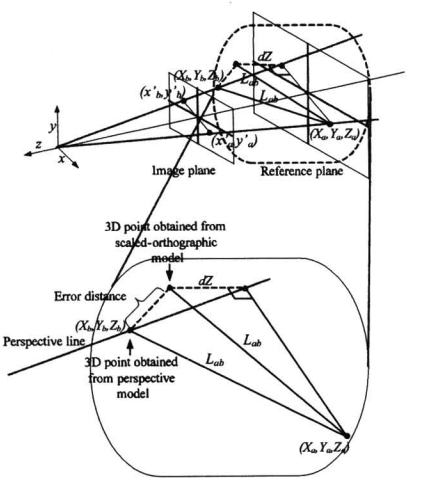


Fig. 1 The scaled-orthographic camera model of two pairs of corresponding points.

approach is a well-known technique proposed by Taylor[7] and used widely in many research works [8]–[10]. It does not require knowledge of camera parameters. In this approach, the basic assumption is that most human parts lie parallel to the image plane. The scaled-orthographic model is shown in Figure 1. By assuming that the actual length of a segment ab, L_{ab}, is known, we can write

$$L_{ab}{}^2 = (X_a - X_b)^2 + (Y_a - Y_b)^2 + (Z_a - Z_b)^2, \quad (1)$$

where (X_a, Y_a, Z_a) and (X_b, Y_b, Z_b) are real-world coordinates of points a and b, respectively. The segment is projected onto the image plane by a scaling factor, s; therefore,

$$\begin{aligned}(x'_a - x'_b) &= s(X_a - X_b)\\ (y'_a - y'_b) &= s(Y_a - Y_b),\end{aligned} \quad (2)$$

where (x'_a, y'_a) and (x'_b, y'_b) are two corresponding points, on the image plane, of points a and b, respectively. By assuming that the actual point is projected perpendicularly to the reference plane, the relative depth between the two points is defined as

$$dZ = (Z_a - Z_b) = \sqrt{L_{ab}{}^2 - \frac{(x'_a - x'_b)^2 + (y'_a - y'_b)^2}{s^2}} \quad (3)$$

Because dZ cannot be complex,

$$s \geq \frac{\sqrt{(x'_a - x'_b)^2 + (y'_a - y'_b)^2}}{L_{ab}}. \quad (4)$$

To obtain the most accurate 3D position from the model, the optimal scaling factor is searched using a comprehensive grid-based optimization. Once the real-world depth of the first point, Z_a, is defined, the depth of the next point, Z_b, is computed using Eq. (1) and Eq. (2). In the reconstruction of 3D articulated human by the concept of this model, the corresponding points on the image plane and the length of each segment are assumed to be known a priori. This method is simple to establish a 3D pose; however, it may not yield accurate results. It is most suitable for telecentric cameras used in most industrial inspection applications but not for the perspective cameras found in general use. An example of the error distance of the 3D reconstruction from the scaled-orthographic model is shown in Figure 1. It is significantly increased with the perspective effects.

To overcome the limitations of the scaled-orthographic approach, Remondino et al. proposed a method based on a perspective concept [16]. In their work, some camera parameters such as focal length must be known a priori. Moreover, at least one predefined segment is assumed to lie on a plane parallel to the image plane. By assuming that the scaling factor is $\frac{F}{Z}$ where F is the principal distance (approximated by the focal length of the lens). The relationships of points on the real-world and the image plane can be represented by

$$\begin{aligned}s_a X_a &= x'_a & s_b X_b &= x'_b\\ s_a Y_a &= y'_a & s_b Y_b &= y'_b,\end{aligned} \quad (5)$$

where (X_a, Y_a, Z_a) and (X_b, Y_b, Z_b) are projected onto the image plane at points (x'_a, y'_a) and (x'_b, y'_b) by the scaling factors s_a and s_b, respectively. Assuming that L_{ab} and F are known,

$$L_{ab}{}^2 = (\frac{x'_a Z_a - x'_b Z_b}{F})^2 + (\frac{y'_a Z_a - y'_b Z_b}{F})^2 + (Z_a - Z_b)^2. \quad (6)$$

From the assumption that at least one segment lies on a plane parallel to the image plane ($Z_a = Z_b$), Z_b value

of that segment can be solved by

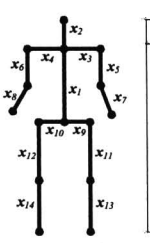
$$Z_b = \frac{FL_{ab}}{\sqrt{x'^2_a - 2x'_a x'_b + x'^2_b + y'^2_a - 2y'_a y'_b + y'^2_b}}. \quad (7)$$

After the depth of the joints on the image plane is computed by Eq. (7), the depth of the next point of the next segment can be computed using Eq. (6). In general, this approach is more accurate than the scaled-orthographic approach. However, the principle length and the length of each segment are assumed to be known a priori. This model also needs an assumption that at least one predefined segment lies parallel to the image plane. The latter assumption is difficult to meet in most general image sequences.

The Taylor method [7] was initially designed for the reconstruction of a single image. However, several researchers [8] have extended the work to 3D reconstruction and tracking. Moreover, the lifting of 2D correspondence to 3D model still provides non-unique solutions. In their methods, the solution for the current frame is selected as the closest configuration to the solution for the previous frame. From this sense, it may diverge from the optimal configuration due to the hard decision in selecting false solutions, especially in arbitrarily complicated poses.

2. Our Proposed 3D Reconstruction

In our approach, the construction of the 3D body is performed from known 2D correspondences based on a 3D skeleton model. Since the relative 3D pose can be constructed at any scale, we therefore define a reference distance as a virtual distance (at the image plane) which the root-node joint lies in the first frame. Firstly, the length of each segment is determined and then the reference distance is computed for the first frame. These parameters are used in the reconstruction of the 3D body in the subsequent frames.



Body parts	Relative length
Head	1.25 T
Torso	2.50 T
Shoulder	2.00 T
Upper arm	1.50 T
Lower arm	2.00 T
Pelvic	1.50 T
Upper leg	2.00 T
Lower leg	2.00 T

Fig. 2 The 3D skeleton human model in our approach.

As shown in Figure 2, the 3D human skeleton in our work is modeled by 15 joint points and 14 segments. The joints are (in order): abdomen, neck, head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle. The skeleton model can be expressed by a tree structure with a root node. Note that in our work, the abdomen joint is selected as the root, since it is relatively more stable than other joints. We also define a relative length of each part based on concept of Leonardo Da Vinci [16], i.e. each part is up to a scale T, as shown in Figure 2.

In case the human subject does not satisfy the Da Vinci's model, errors of segment lengths directly effect the joint position computation, especially at the end joints due to error accumulation. The severity depends upon the degree of the mismatch. The error can be alleviated by considering corresponding symmetric parts and introducing relaxation into the reconstruction.

2.1 Problem formulation

Our approach for the 3D reconstruction of a human pose from known coordinates of a 2D-point correspondence is based on a perspective camera concept, as shown in Figure 3. The 2D points obtained from an uncalibrated camera serve as input for our work.

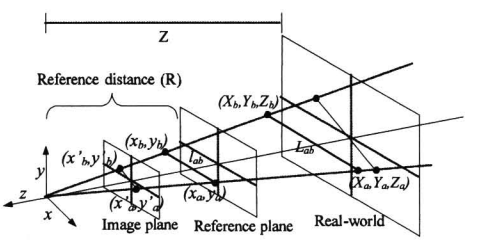


Fig. 3 The perspective camera model of two pairs of corresponding points.

In Figure 3, (X_a, Y_a, Z_a) and (X_b, Y_b, Z_b) are real-world coordinates of points a and b, respectively. $v_a = (x_a, y_a, z_a)$ and $v_b = (x_b, y_b, z_b)$ are corresponding relative coordinates of the points a and b around the reference distance, respectively. By the word "around the reference distance", we mean both 3D points are scaled using the same magnification $(\frac{R}{Z})$ where Z is the real world depth of the root node joint at the first frame and R is the reference distance from the origin to the image plane. (x'_a, y'_a) and (x'_b, y'_b) are the respective image coordinates of points a and b at the image plane. L_{ab} and l_{ab} are the lengths between points a and b in terms of real-world distance and virtual distance around the reference distance, respectively. The magnification of a coordinate around the reference distance to the real-world coordinate is represented by

$$\frac{1}{\alpha} = \frac{x_a}{X_a} = \frac{y_a}{Y_a} = \frac{z_a}{Z_a} = \frac{l_{ab}}{L_{ab}} = \frac{R}{Z}, \quad (8)$$

From Figure 3, we can project the points v_a and v_b onto the image plane and obtain the following relationships:

$$\begin{array}{llll} x_a = s_a x'_a & X_a = \alpha s_a x'_a & x_b = s_b x'_b & X_b = \alpha s_b x'_b \\ y_a = s_a y'_a & Y_a = \alpha s_a y'_a & y_b = s_b y'_b & Y_b = \alpha s_b y'_b \\ z_a = s_a R & Z_a = \alpha s_a R & z_b = s_b R & Z_b = \alpha s_b R, \end{array} \quad (9)$$

where s_a and s_b are scaling factors corresponding to the ratios of points Z_a and Z_b to R, respectively.

v_a and v_b are up-to-scale coordinates of the corresponding 3D real-world coordinates. To estimate 3D real-world coordinates, knowledge of either a 3D real-world coordinate or an actual segment length is required by Eq. (9). However, up-to-scale coordinates are adequate in most character-pose-related applications.

2.2 Proposed 3D reconstruction

In the construction of 3D relative coordinates of the body, we can represent a human skeleton model as a tree structure $\{\mathbf{v}, \mathbf{E}\}$. The nodes are represented by $\mathbf{v} = \{v_i | i \in [1, ..., N]\}$, where v_i is the 3D relative coordinate of the i^{th} joint, and N is the total number of joints. The edge or connection between two end nodes of a link is denoted by $\varepsilon \in \mathbf{E}$, where $\mathbf{E} = \{\varepsilon_{ab} | a, b \in \mathbf{v}\}$. The length of ε_{ab} is represented by l_{ab}. Each joint is projected onto the image plane by a scaling factor.

2.2.1 Determining reference distance

In our approach, the reference distance is required and is determined in the first frame. We set an assumption that a set of three known joints forming a right triangle with only one leg lain on a plane parallel to the image plane in the first frame. This can be easily achieved by a human stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane. An example of the human pose satisfying the assumption is shown in Figure 4. First, we find the reference distance by considering a segment that is not in a plane parallel to the image plane and denoted it by ε_{qr} with length l_{qr}, as shown in Figure 4. For segment selection, a segment with right angle with respect to the vertical axis of the human body is chosen. In Figure 4, $v_p = (x_p, y_p, z_p)$, $v_q = (x_q, y_q, z_q)$ and $v_r = (x_r, y_r, z_r)$ are virtual coordinates of segment end points around the reference distance. l_{pq} is parallel to the image plane and can be determined using Euclidean distance of its endpoint image coordinates. l_{qr} is proportional to l_{pq} according to the Da Vinci's ratio. For the reference distance computation, the projected end points of the link are used to obtain

$${l_{qr}}^2 = (s_q x'_q - s_r x'_r)^2 + (s_q y'_q - s_r y'_r)^2 + (s_q R - s_r R)^2.$$

Hence, the reference distance can be described by the following equation:

$$R = \sqrt{\frac{{l_{qr}}^2 - (s_q x'_q - s_r x'_r)^2 - (s_q y'_q - s_r y'_r)^2}{(s_q - s_r)^2}}, \quad (10)$$

where (x'_q, y'_q) and (x'_r, y'_r) are image coordinates of points q and r, respectively; s_q and s_r are the respective corresponding scaling factors of points q and r; and l_{qr} is its length.

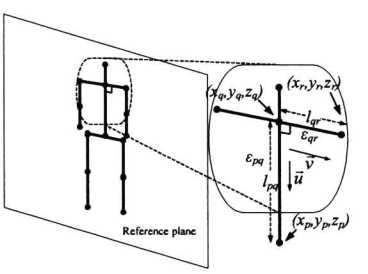


Fig. 4 Initial posture in the first frame that two joints are not in the same plane.

From Eq. (10), s_q and s_r must be calculated in the determination of the reference distance. From the assumption made in the first frame, the axis of the scaled human body is set to lie on the image plane such that scaling factors regarding the abdomen and neck joints are set to unity. In the determination of image distance, we select the link between the neck and the abdomen joints as vector $\vec{u}$ and the link between the neck and the left shoulder joints as vector $\vec{v}$. The scaling factor of the abdomen and neck joints is 1 and $\vec{u}$ is perpendicular to $\vec{v}$, i.e., $\vec{u}.\vec{v} = 0$ (only required for the first frame), and hence

$$\begin{aligned} \vec{u} &= (x'_p - x'_q, y'_p - y'_q, 0) \\ \vec{v} &= (s_r x'_r - x'_q, s_r y'_r - y'_q, s_r R - R) \end{aligned}$$

and

$$s_r = \frac{x'_p x'_q - {x'_q}^2 + y'_p y'_q - {y'_q}^2}{x'_p x'_r - x'_q x'_r + y'_p y'_r - y'_q y'_r}. \quad (11)$$

(x_p, y_p, z_p), (x_q, y_q, z_q) and (x_r, y_r, z_r) are the abdomen, neck and left shoulder joints around the reference distance and are projected onto the image plane at points (x'_p, y'_p), (x'_q, y'_q) and (x'_r, y'_r) by scaling factor s_p, s_q and s_r, respectively. By putting $s_q = 1$ and s_r obtained from Eq. (11) back to Eq. (10), the reference distance is determined.

2.2.2 Determining scaling factor of root-node

The position of the root segment is very important, because it is the starting point for the reconstruction of the 3D body in our approach. From our assumption,

the scaling factor of the root-node joint, s_1 is unity at the first frame. However, it changes and, therefore, requires computation for subsequence frames due to human motion. It is determined by a comprehensive grid-based search with the constraint that its value cannot be negative. We can also perform a faster search for s_1 within a small range around s_1 of a previous frame. The values satisfying 3D human pose based on a joint-angle limitation constraint (Section 3.1) are selected as s_1 in the current frame.

2.2.3 Determining scaling factor of other nodes

For the reconstruction and track 3D relative positions of articulated body, the scaling factor of each joint is computed. Suppose s_a is known from result of the previous computed segment, s_b can be determined by finding intersection points between a sphere and a line. The starting joint $(s_a x'_a, s_a y'_a, s_a R)$ and length l_{ab} of segment ε_{ab} are served as the center and the radius of the sphere, respectively. By assuming that, we can write

$$(x_b - s_a x'_a)^2 + (y_b - s_a y'_a)^2 + (z_b - s_a R)^2 = {l_{ab}}^2. \quad (12)$$

The line started from the origin $(0,0,0)$ pointing in the direction (x'_b, y'_b, R) meets the sphere at point $(x_b, y_b, z_b) = (s_b x'_b, s_b y'_b, s_b R)$.

Since (x'_a, y'_a) and (x'_b, y'_b) are data points, R and l_{ab} are computed from the first frame, only s_b remains to be calculated. We also perform the determination for $s_k, k = 2, ..., N$ sequentially using, s_a obtained from the result of the previous computed segment. By simple manipulation, Eq. (12) becomes

$$({x'_b}^2 + {y'_b}^2 + R^2){s_b}^2 - (2s_a x'_a x'_b + 2s_a y'_a y'_b + 2s_a R^2)s_b + ({s_a}^2{x'_a}^2 + {s_a}^2{y'_a}^2 + {s_a}^2 R^2 - {l_{ab}}^2) = 0.$$

Therefore,

$$s_b = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}, \quad (13)$$

where

$$A = {x'_b}^2 + {y'_b}^2 + R^2$$
$$B = -(2s_a x'_a x'_b + 2s_a y'_a y'_b + 2s_a R^2)$$
$$C = {s_a}^2{x'_a}^2 + {s_a}^2{y'_a}^2 + {s_a}^2 R^2 - {l_{ab}}^2.$$

It can be seen that the solutions are not unique. This problem is inherent to all lift-up approaches. To select the best configuration, some further assumptions, e.g. that all joints lie on the same plane [7] are normally assumed. For our technique, we delay the selection of each best configuration of the first two frames to the third frame. This makes our method even less restricted.

3. Our Proposed 3D Body Tracking

After the first frame, 3D reconstruction can be more conveniently performed using tracking. At each frame, multiple configurations of 3D reconstructions are obtained. To find the optimal solution, we applied a joint angle limitation constraint in order to prune spurious solutions and to avoid singularity. Moreover, we employed a technique based on the concept of multiple hypothesis tracking (MHT) [13] with a motion-smoothness function to select the best solution based on temporal information.

3.1 The joint-angle limitation constraint

Multiple configurations per frame are obtained due to non-uniqueness of the reconstruction. Some configurations are singular or spurious solutions. To remove some of these undesired configurations, we employ a joint-angle limitation that limits arm and leg angles to fall between 10 and 180 degrees, as shown in Figure 5.

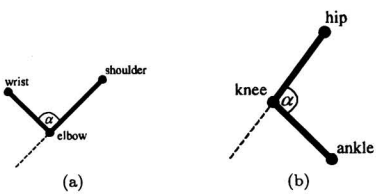


Fig. 5 Joint-angle limitation: (a) Arm (b) Leg.

3.2 Optimal solution selection by MHT

Even with the constraint specified, there are still a great number of possible configurations left. Techniques using temporal information have been introduced to select the optimal solution in other works. Most of these approaches [16] are based on choosing the closest configuration from one frame to the next. This method only maintains the solution for the last frame. It works well in some simple poses but fails in a larger number of cases; it can also lead to divergences due to an incorrect decision in early frames. We therefore proposed an efficient technique based on MHT to select the best configuration at the current frame using the past temporal information while still maintaining a number of most-likely previous poses.

We define the *temporal joint trajectory* of a joint as a time sequence of the distances from the 3D relative position of that particular joint to the coordinate origin, i.e., the camera center. The smoothness of the trajectory is used to evaluate the fitness of a hypothesis instead of the hypothesis probability used in the original study [13]. A polynomial function and least squares method can be used to develop a motion model as well as an estimator in the context of MHT. In our experiment (not shown here), we found that a simple line

connecting two previous frames of the joint sequence was adequate and very efficient, and therefore, was used it in our work. The details on the formulation of the technique are described below.

At frame t, a pose configuration is denoted by $\theta_i(t)$, and the set of all possible configurations is represented by $\{\theta_i(t)\}$, $i = 1, 2, ..., N_t$. From MHT concept, a hypothesis at frame t, which is denoted by Θ_i^t, is a pose sequence and can be written as

$$\Theta_i^t = \{\Theta_{m(i)}^{t-1}, \theta_i(t)\}, \tag{14}$$

where $\Theta_{m(i)}^{t-1}$ is the parent pose sequence at frame $t-1$ of Θ_i^t. To determine the fitness of a hypothesis, the smoothness of the trajectory is used and computed by

$$E_i^t = \Delta(\theta_i(t)) + E_{m(i)}^{t-1}, \tag{15}$$

where

$$\Delta(\theta_i(t)) = \sum_{j=1}^{N} \frac{|t + A_{i,j}^t D_{i,j}^t + B_{i,j}^t|}{\sqrt{A_{i,j}^{t\ 2} + 1}}. \tag{16}$$

E_i^t is the fitness of the pose sequence i at frame t, and $D_{i,j}^t$ is the distance from joint j of the configuration i at frame t to the origin. $A_{i,j}^t$ and $B_{i,j}^t$ are parameters of the line equation passing through the points $(t-2, D_{i,j}^{t-2})$ and $(t-1, D_{i,j}^{t-1})$, respectively. N is the number of joints in the human body model.

The fitness value is calculated for each of the possible pose sequences in each hypothesis of MHT. The fitness values in each hypothesis of MHT are then ordered. The configuration corresponding to the best fitness is selected as the solution for that frame, while k best configurations according to their fitness values are kept for generating new hypotheses of MHT for the next frame. The computational complexity is $O(2^N k \log(2^N k))$ by mean of conventional tree sorting implementation in the MHT. Figure 6 shows a flowchart of the algorithm.

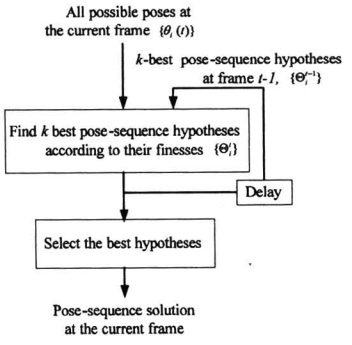


Fig. 6 Flowchart of the multiple hypothesis algorithm.

Our method bears similarity to Howe's work [6] in the sense that more temporal information than the previous frame is used. However, Howe's method performs a comprehensive search of all possible pose sequences in the past few frames, while our method requires a search of only the k best possible pose sequences in the past few frames. In addition, Howe's 3D position configurations per frame are obtained by consulting an image feature database; thus, the number of configurations obtained may be different from ours.

4. Experiments

To evaluate the performance of our proposed method, we tested our approach on both synthesized and real-world image sequences. We assume that a human stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane, and that all segments extracted from a monocular view are available in the first frame. To benchmark the performance of our proposed method, we compare our approach with both scaled-orthographic [7] and previous perspective approaches [16]. In all of our experiments, we selected the link between the neck and left shoulder of the human for the reference distance computation due to its perpendicular angle assumption as well as the abdomen joint for the root joint due to its relatively small movements. However, we found (from our experiments not shown here) that the selection of other, less stable joints as the root joint only slightly affected the performance of the result. To evaluate the performance of reconstruction of 3D articulated body, 3D pose results are scaled to best match and then compared with 3D real world coordinates of the ground-truth.

4.1 Performance comparison with other approaches

We synthesized four 3D human image sequences in order to compare our proposed method with both scaled-orthographic [7] and previous perspective [16] approaches in terms of accuracy. The 2D projected joints from those 3D synthesized data are served as the input in the comparison. The image sequences include aerobic style activities, which contain 15 frames in each scene. Attributes of each scene are shown in Table 1. Since the orthographic method is originally designed for 3D reconstruction from a single image, it is not fair to perform benchmarking based on tracking results due to the accumulated errors in the latter frames of the sequences. Therefore, we exclude solving an ill-posed problem by selecting the closest configuration to the ground truth for all candidate algorithms at each frame.

Some ground truth frames of scenes 1-4 are shown in the first column of Figures 7 (a)-(d). Examples of results from our approach are shown in the second col-

Table 1 Attributes of four test image sequences.

Scene	Magnitude of Perspective Effect	At Least One Segment Lies on a Parallel Plane
Scene 1	Weak	Yes
Scene 2	Weak	No
Scene 3	Strong	Yes
Scene 4	Strong	No

umn. Similar results from the scaled-orthographic and previous perspective approaches are presented in the last two columns, respectively.

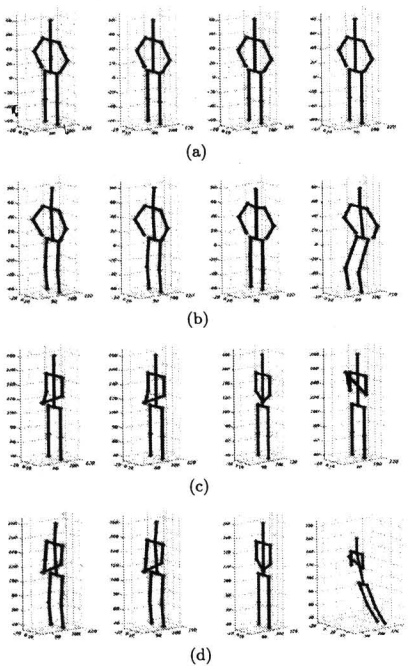


Fig. 7 Samples of ground truth and samples of results from our approach, Taylor and Remondino techniques in (a) scene 1 (b) scene 2 (c) scene 3 and (b) scene 4.

From the experimental results, the averaged per-joint error distances from the ground truth for our proposed method, the scaled-orthographic approach [7] and the previous perspective [16] approach are shown in Table 2. It can be seen that the results of our approach are more accurate than other approaches for all sequences. The scaled-orthographic approach [7] performed very well in the first sequence because the scene contained a weak perspective effect and relatively simple poses. For the previous perspective approach [16], the results were very good and comparable to ours in scenes 1 and 3; this is because each frame of those sequences has at least one segment in a parallel plane. Note that our approach outperforms the others for sequences with both weak and strong perspective effects.

Table 2 Averaged per-joint error distances (in centimeters) from different approaches.

Scene	Our Approach	Scaled-orthographic Approach [7]	Previous Perspective Approach [16]
Scene 1	0.00	1.60	0.08
Scene 2	0.10	2.34	7.21
Scene 3	0.90	4.25	1.01
Scene 4	0.89	4.92	9.68

4.2 Results on real-world images

In addition, we tested our approach using three real-world image sequences†. We used the known 2D position of each joint obtained manually in the image as input data for our proposed method. The real-word image sequences include walking, ball-throwing and back-flipping in a Mocap dataset [14] containing 300, 800 and 239 frames, respectively. Some results of the walking, ball throwing and back flipping scenes from our approach are shown in Figures 8(a), (b) and (c), respectively. In this experiments, the 1000 best hypotheses are maintained at each frame by our tracking method.

The results show that the accuracy of our proposed method is very high in all sequences. On average, the error distance of the walking, ball-throwing and back-flipping scenes as compared with the corresponding ground truth is 1.25, 1.21 and 1.97 centimeters, respectively.

4.3 Performance of 3D tracking with motion-smoothness function

All trajectories of the left wrist in the walking, ball-throwing and back-flipping scenes are shown by dot marker in Figures 9(a), (b) and (c), respectively. It can be seen that multiple trajectories were obtained. This complicates the selection of the optimal solution. The trajectories of the left wrist after applying the joint-angle-limit constraint are shown by plus marker in Figures 9(a), (b) and (c), respectively.

It can be seen that our approach shows accurate results for both simple and strong perspective-effect cases. Moreover, the ground truth is always found (with very little uncertainty) within the set of configurations computed from our reconstruction of the 3D human pose. MHT can identify the optimal solution for all images in each image sequence as shown by overlapped parts between solid lines and square makers in Figures

†Since the sequences do not meet assumptions made by the other two approaches, the results from those techniques were very inaccurate and not shown here to save space.

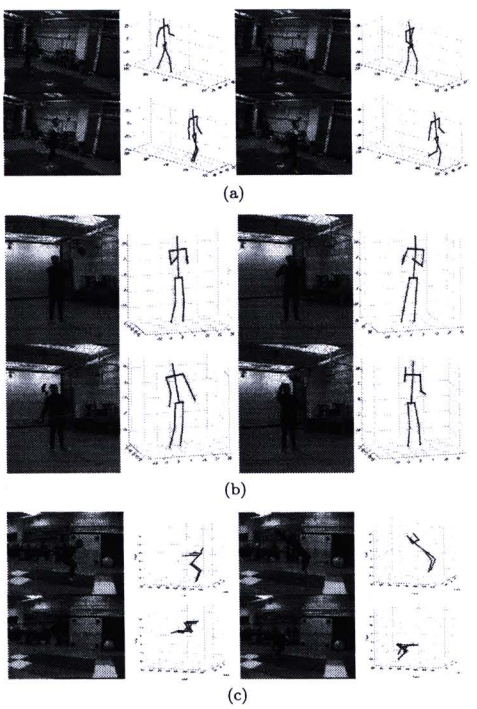


Fig. 8 Some results for the real-world images in (a) walking, (b) ball throwing and (c) back flip sequences.

9(a), (b) and (c), respectively.

4.4 Robustness against errors in reference distance determination

In this section, we investigated robustness of our approach due to errors in the reference distance estimation. It is computed in the first frame and used in the other frames. In our experiment, the reference distance R is varied between 70% to 130%. The average distance errors from the corresponding ground truth are shown in Figure 10. The average distance errors are between 1 to 6 centimeters per joint. By comparing the error to the human height of approximately 175 centimeters, the method is rather robust against errors in the reference distance determination.

4.5 Robustness against noise in 2D data

Additionally, we investigated robustness of our proposed method on noise in the 2D data. We used synthesized data of 2D joint locations corrupted by a Gaussian noise. 2D Gaussian distributions with zero mean and variances 0.5, 1, 1.5, 2, 2.5 and 3 were generated and

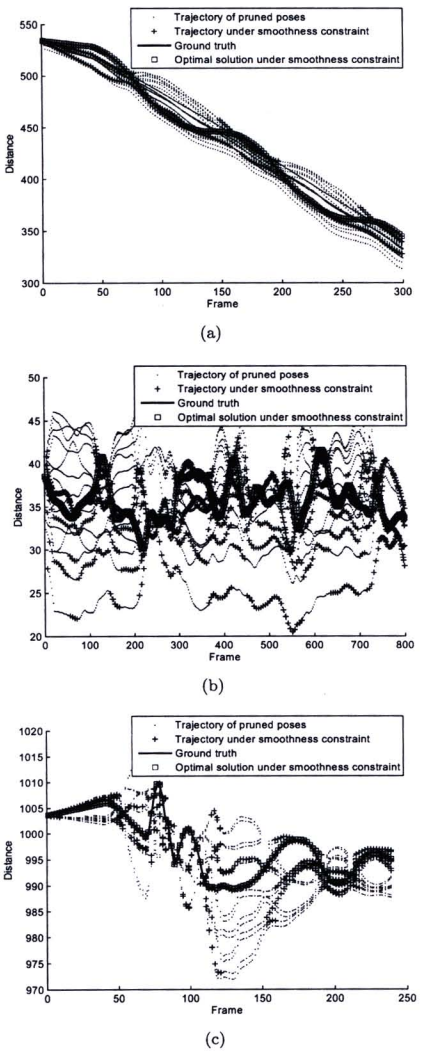


Fig. 9 Trajectories of the left wrist after tracking in (a) walking, (b) ball-throwing and (b) back-flipping sequences.

used in the experiments. The average distance errors from the corresponding ground truth are 2.15, 3.00, 2.15, 2.98, 3.84 and 3.94 centimeters, respectively. It can be seen that our approach is robust against variation in the 2D input data.

5. Conclusions

A novel method is presented to estimate 3D positions of

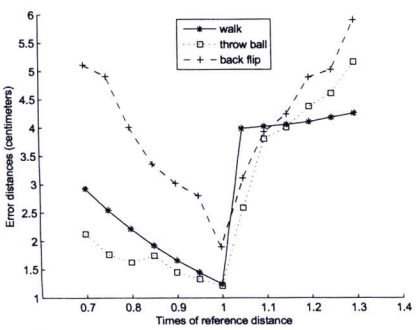


Fig. 10 Error distance due to errors in difference reference distance determination

an articulated body from 2D point correspondences in a single uncalibrated image. It is based on the perspective concept. However, it does not require any camera parameters from the user. Instead, our method computes a reference distance under a simpler assumption in the first frame. A closed-form solution is provided for 3D reconstruction. However, multiple configurations are obtained from the method because the problem is inherently non-uniqueness. An MHT-based tracking technique is introduced in order to select the best solution. This method is not only effective but also very efficient. Our approach was tested on both synthesized and real-world image sequences, and we obtained excellent results. Quantitative comparisons with scaled-orthographic [7] and previous perspective [16] approaches with respect to accuracy were also performed using image sequences with different degrees of perspective and pose complexities. Our proposed method outperformed these other approaches [7], [16], especially in scenes with strong perspective effects and various poses that are generally difficult to model.

Acknowledgement

We gratefully acknowledge the support from the Nation University Research (NRU) program as well as Office of the Higher Education Commission (OHEC).

References

[1] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H. Seidel, Markerless motion capture with unsynchronized moving cameras, In Proc. CVPR, 2009.

[2] J. Gall, C. Stoll, E. Aguiar, C. Theobalt, B. Rosenhahn, and H. Seidel, Motion capture using joint skeleton tracking and Surface Estimation, In Proc. CVPR, 2009.

[3] C. Barron, and I.A. Kakadiaris, Estimating anthropometry and pose from a single image, CVIU, 2001. p.269-284.

[4] F. Yang, and X. Yuan, Human movement reconstruction from video shot by a single stationary camera, Journal of Annals of Biomedical Engineering, 2005(33). p.674-684.

[5] Z. Jianhui, L. Li, and K.C. Keong, 3d posture reconstruction and human animation from 2d feature points, Computer Graphics Forum, 2005(4). p.759-771.

[6] N.R. Howe, Silhouette lookup for monocular 3d pose tracking, Image Vision Computing, 2007(14). p.331-341.

[7] C.J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, In Proc. CVPR, 2000. p.677-684.

[8] G. Mori, and J. Malik, Recovering 3D human body configurations using shape contexts, TPAMI, 2006(28). p.1052- 1062.

[9] W. Lao, J. Han, and P.H.N. de With, 3D modeling for capturing human motion from monocular video, In Proc. Symposium on Information Theory in the Benelux, 2006.

[10] I. Rius, D. Rowe, J. Gonzlez, and F.X. Roca, 3D Action modeling and reconstruction for 2D human body tracking, In Proc. CVPR, 2005. p.146-154.

[11] B. Zou, S. Chen, C. Shi, and U.M. Providence, Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking, Pattern Recognition, 2009(42). p. 1559-1571.

[12] X. Peng, B. Zou, S. Chen, and P. Luo, Reconstruction of 3D human motion pose from uncalibrated monocular video sequence, Int.J. of Information and Systems Sciences, 2009(5). p.503-515.

[13] I.J. Cox, and S.L. Hingorani, An Efficient implementation of Reid's multiple hypothesis tracking algorithm and Its evaluation for the purpose of virtual tracking, TPAMI, 1996(18).

[14] Mocap dataset. http://mocap.cs.cmu.edu/motcat.php

[15] L. Zhang, and L. Li, Human animation from 2D correspondence based on motion trend prediction Source, In Proc. WSEAS Artificial Intelligence, 2006.

[16] F. Remondino, and A. Roditakis, Human figure reconstruction and modeling from single image or monocular video sequence, In Proc. 3D Digital Imaging and Modeling, 2003.

Kittiya Khongkraphan received the B.Sc. degree in Mathematics and M.Sc. degree in Computer Science from Prince of Songkla University, Thailand, in 1991 and 2000, respectively. Currently, she is a Ph.D. candidate of King Mongkut's University of Technology Thonburi (KMUTT), Thailand. Her research interests are computer vision and image processing.

Pakorn Kaewtrakulpong received the B.Eng. degree in Electrical Engineering from KMUTT in 1992. He received M.Sc. and Ph.D. degrees in Systems Engineering from Brunel University in 1998 and 2002, respectively. He is currently an associate professor at the faculty of engineering, KMUTT. His research interests include machine vision, industrial automation and instrumentations.

CURRICULUM VITAE

NAME	Mrs. Kittiya Khongkraphan
DATE OF BIRTH	12 October 1969
EDUCATION RECORD	
BACHELOR'S DEGREE	Bachelor of Science (Mathematics), Prince of Songkla University, 1991
MASTER'S DEGREE	Master of Science (Computer Science), Prince of Songkla University, 2000
DOCTORAL DEGREE	Doctor of Philosophy (Electrical and Computer Engineering) King Mongkut's University of Technology Thunburi, 2011
SCHOLARSHIP/RESEARCH GRANT	Ph.D. Scholarship from Commission on Higher Education, Ministry of Education, Thailand
EMPLOYMENT RECORD	Lecturer Department of Mathematics and Computer Science Faculty of Science and Technology Prince of Songkla University
PUBLICATIONS	Khongkraphan, K. and Kaewtrakulpong, P., 2007, "Robust Contour Tracking in Cluttered Background using Snake on Weighted- Gradient Image", **Proceedings International Workshop on Advanced Image Technology**, 8-9 January 2007, Bangkok, Thailand, pp. 638-642. Khongkraphan, K. and Kaewtrakulpong, P., 2010, "A Novel Method of 2D Articulated Body Tracking under Self-occlusion and Ambiguity", **IEICE Electronics Express (ELEX)**, Vol. 7 (2010), No. 15, pp. 1106-1111. Khongkraphan, K. and Kaewtrakulpong, P., 2011, "Efficient Human Body Tracking by Quick Shift Belief Propagation", **IEICE Transaction on Information & Systems**, Vol. E94-D, No. 4, pp. 905-912.

Khongkraphan, K. and Kaewtrakulpong, P., 2011, "A Novel Reconstruction and Tracking of 3D-articulated Human Body from 2D Point Correspondences of a Monocular Image Sequence", **IEICE Transaction on Information & Systems**, Vol. E94-D, No. 5.

King Mongkut's University of Technology Thonburi

Agreement on Intellectual Property Rights Transfer for Postgraduate Students

Date **1 March 2011**

Name **Mrs. Kittiya** Middle Name..... Surname/Family Name **Khongkraphan**
Student Number **48530001** who is a student of King's Mongkut's University of Technology Thonburi (KMUTT) in ❍ Graduate Diploma ❍ Master Degree ☑ Doctoral Degree
Program **International program** Field of Study **Computer Engineering**
Faculty/School **Faculty of Engineering**
Home Address **365/343 Condobansuangton Bangmod Thungkru Bangkok**
Postal Code **10140** Country **THAILAND**

I, as 'Transferer', hereby transfer the ownership of my thesis copyright to King's Mongkut's University of Technology Thonburi who has appointed (Dean's name) **Assoc. Prof. Dr. Piuabutr Wanichpongpan** Associate Dean for Academic Affairs (Acting for Dean) to be 'Transferee' of copyright ownership under the 'Agreement' as follows.

1. I am the author of the thesis entitled **Efficient 3D Pose Estimation and Tracking of Articulated Body from Uncalibrated Monocular Image Sequence** under the supervision of **Assoc. Prof. Dr. Pakorn Kaewtrakulpong** who is my supervisor in accordance with the Thai Copyright Act B.E. 2537. The thesis is a part of the curriculum of KMUTT.

2. I hereby transfer the copyright ownership of all my works in the thesis to KMUTT throughout the copyright protection period in accordance with the Thai Copyright Act B.E. 2537, effective on the approval date of thesis proposal consented by KMUTT.

3. To have the thesis distributed in any form of media, I shall in each and every case stipulate the thesis as the work of KMUTT.

4. For my own distribution of thesis or the reproduction, adjustment, or distribution of thesis by the third party in accordance with the Thai Copyright Act B.E. 2537 with remuneration in return, I am subject to obtain a prior written permission from KMUTT.

5. To use any information from my thesis to make an invention or create any intellectual property works within ten (10) years from the date of signing this Agreement, I am subject to obtain prior written permission from KMUTT, and KMUTT is entitled to have intellectual property rights on such inventions or intellectual property works, including entitling to take royalty from licensing together with the distribution of any benefit deriving partly or wholly from the works in the future, conforming with the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.

6. If the benefits arise from my thesis or my intellectual property works owned by KMUTT, I shall be entitled to gain the benefits according to the allocation rate stated in the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.

Signature.....Transferor

(Mrs. Kittiya Khongkraphan)

Signature.....Transferee

(Assoc. Prof. Dr. Piyabutr Wanichpongpan)

Associate Dean for Academic Affairs (Acting for Dean)

Signature.....Witness

(Assoc. Prof. Dr. Tiranee Achalakul)

Signature.....Witness

(Assoc. Prof. Dr. Pakorn Kaewtrakulpong)